Incorporating spatial autocorrelation and association in the statistical null model test of co-occurrence

by

Vitalis Kimutai Lagat

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Mathematics in the Faculty of Science at Stellenbosch University

Department of Mathematical Sciences,
Mathematics Division,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisor: Prof. Cang Hui

March 2017

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: Vitalis Kimutai Lagat .....

Date: March 2017 .....

Abstract

Incorporating spatial autocorrelation and association in the statistical null model test of co-occurrence

Vitalis Kimutai Lagat

Department of Mathematical Sciences,
Mathematics Division,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: Masters by Research (MRes)

December 2016

To avoid conflicts and optimally exploit environmental resources, species will partition available habitats, forming co-occurrence patterns. Such datasets are often described as a species-by-site matrix. Null models based on permutations with constraints on row or column sums have been used in this regard, with the Chessboard score (C-score) a common metric for detecting significant signals of association or dissociation, from which the type of biotic interactions can be inferred. However, such a permutation test often ignore the spatial autocorrelation of species distributions which could lead to counterintuitive results in the null model test. Consequently, tests should account for the spatial autocorrelation of each species. Another important concept that is ignored in the classic permutation test is the matching of environmental heterogeneity and species' habitat preference. To tease apart the role of environmental heterogeneity from biotic interactions, the permutation test should also be allowed to reserve the association between species. This project thus designs a permutation null model test that can progressively include the spatial autocorrelation of species and the association between species so that the role of aggregation and environmental heterogeneity can be further examined. A R package has been designed to implement both classic (spatially implicit) null model tests of co-occurrence and newly designed approaches for the permutation test with constraints on species autocorrelation and association. Though both the classic and the newly designed null models lead to the same inference regarding inter-specific competition as a factor structuring ecological communities, the latter is more reliable because it does not violate any of the assumptions of

the test.

Keywords: Null model; interspecific competition; spatial autocorrelation; species association; species co-occurrence; null hypothesis; species-by-site matrix; permutation test; checkerboard distribution.

Uittreksel

Inkorporeer ruimtelike outokorrelasie en assosiasie in die statistiese nul model toets van mede-voorkoms

Vitalis Kimutai Lagat

Departement Wiskundige Wetenskappe,
Afdeling Wiskunde,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: Meesters deur navorsing (MRes)

Desember 2016

Om con IKT omgewingshulpbronne te vermy en optimaal te benut, spesies verdeling beskikbaar habitatte, die vorming van mede-voorkoms patrone. Sulke datastelle is dikwels beskryf as 'n spesie-deur-site matriks. Nulmodelle gebaseer op permutasies met beperkings op ry of kolom bedrae is gebruik in hierdie verband, met die Skaakbord telling (C-telling) 'n algemene maatstaf vir die opsporing van betekenisvol cant seine van assosiasie of dissosiasie, waaruit die tipe biotiese interaksies kan wees afgeleide. Maar so 'n permutasie toets dikwels ignoreer die ruimtelike outokorrelasie spesies uitkerings wat kan lei tot counter resultate in die nul model toets. Gevolglik moet toetse rekening vir die ruimtelike outokorrelasie van elke spesie. Nog 'n belangrike konsep wat geÃŕgnoreer in die klassieke permutasie toets is die passing van die omgewing heterogeniteit en spesie se habitat voorkeur. Om terg uitmekaar die rol van die omgewing heterogeniteit uit biotiese interaksies, moet die permutasie toets ook toegelaat word om behou die assosiasie tussen spesies. Hierdie projek ontwerp dus 'n permutasie nul model toets wat progressief die ruimtelike outokorrelasie kan insluit spesies en die assosiasie tussen spesies sodat die rol van samevoeging en omgewing heterogeniteit kan verder ondersoek word. A R pakket het is ontwerp om beide klassieke (ruimtelik implisiete) null model toetse te implementer van mede-voorkoms en nuwe ontwerp benaderings vir die toets permutasie met beperkings op spesies outokorrelasie en assosiasie. Alhoewel beide die klassieke en die nuwe ontwerp nulmodelle lei tot dieselfde gevolgtrekking met betrekking tot inter-spesifieke c kompetisie as 'n faktor strukturering ekologiese gemeenskappe, die Laasgenoemde is meer betroubaar

omdat dit nie enige van die aannames van die toets te skend.

Sleutelwoorde: Null model; interspeci kompetisie c; ruimtelike outokorrelasie; spesies assosiasie; spesies mede-voorkoms; nulhipotese; spesie-deur-site matriks; permutasie toets; checker verspreiding.

Acknowledgements

I express my deep and profound gratitude to my supervisor Prof. Cang Hui for his effort, patient guidance and enthusiastic encouragement throughout the course of this research work. The administrative assistance offered by Mrs. Vanessa Du Plessis is much acknowledged. The funding support from AIMS, NRF and the top-up from the SARChI chair Prof. C. Hui is highly appreciated.

Dedications

To my mother (Consolata Jelagat S.) who nurtured me to grow into a hard-working person and inspired me to stand for righteousness in the midst of an evil world.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The beauty of nature never ceases to amaze us. From how organisms organise themselves to optimize on their survival rates, to how ecological communities are structured to form quantifiable and measurable patterns, has not only been a source of merriment through tourism to human kind, but also one of the cornerstones of scientific research.

One such pattern of organisms in ecological communities is the pattern of species co-occurrences. Figure 1.1 illustrates the steps involved in inferring ecological interpretation from these patterns. Scientists have endeavoured to understand the forces behind the co-distribution and co-occurrence patterns of different species in different ecological communities. Some of the questions which have been raised include:

1. What structures ecological communities?
2. Do different species coexist by random chance?
3. Do patterns observed in nature a result of biotic interactions between different species or a random process?

To answer these questions, empirical studies have been carried out on different species to ascertain the forces behind their co-occurrences and generally to elucidate the patterns of their existence on earth. One such study was conducted on the avian communities of the islands of Bismarck Archipelago (Diamond, 1975). Diamond (1975) observed, among other things, that the co-occurrence patterns exhibited by different species of birds are a result of inter-specific competition (refer to Figure 1.2). This means that the avian species who don't compete for the limited resources are more likely to co-occur in the same site, whereas those who compete try as much as possible to avoid each other. This made the presence of certain species on a given site the prerequisite for a successful colonisation by other species.

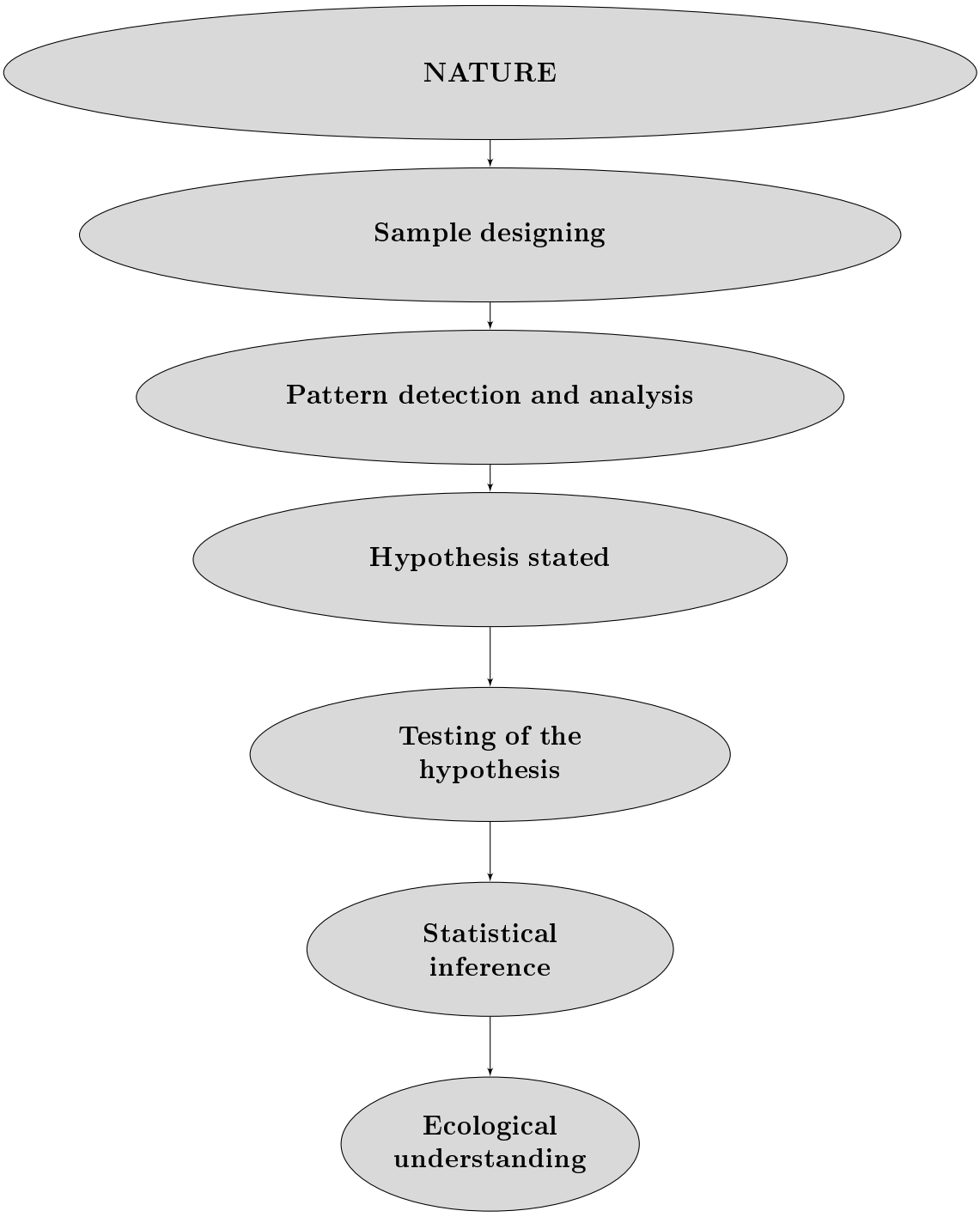


Figure 1.1: Processes involved in quantifying and testing spatial patterns. From these patterns, generalizations and hypotheses can be drawn about the ecological processes. Specific experiments or models can then be used to test the newly defined hypotheses. Some statistical interpretations and ecological understanding can finally be reached. *Source:* (Fortin and Dale, 2005)

(a) Plant interspecific competition

(b) Animal interspecific competition

Figure 1.2: Interspecific competition: 1.2a represents different species of plants competing for sunlight for photosynthesis (PlantLife, 2016). 1.2b represents interspecific competition between animals (Lay, 2016).

Though this was convincing enough, it was later observed that the same patterns alleged by Diamond (1975) to be a result of interspecific competition could be produced by null models (Connor and Simberloff, 1979). This was done by answering the question: how would a community structure look like in the absence of biotic interactions? To answer the question, Monte Carlo simulations were used to generate "null" or artificial communities by randomly reassigning species to islands, which produced artificial communities with similar patterns as those observed by Diamond (1975), without the forces of interspecific competition. This arose a heated debate which has continued unabated to date.

However, though the controversy has persisted, null models have increasingly been viewed as important statistical tools which have been applied to a diverse set of problems in community ecology, giving rise to new insights into mechanisms and patterns of existence of organisms in nature. Further research (Gotelli *et al.*, 1996; Gotelli, 2000, 2001; Gotelli and McCabe, 2002; Krasnov *et al.*, 2006; Hausdorf and Hennig, 2007) have shown that a lot can be depicted from the null models regarding patterns of species co-occurrences.

1.1 Background and Rationale

To attenuate conflicts while utilizing environmental resources, species will partition available habitats forming co-occurrence patterns. Species-by-site matrices have often been used to describe such datasets. In this regard, permutation null model tests with constraints on column or row totals have been employed, with the C-score (checkerboard score) a common metric for detecting significant signals of association or dissociation (Gotelli, 2000; Hui, 2015), from which inference can be made on the type of biotic interactions. However, spatial autocorrelation of the distributions of species have often been ignored by such a permutation null model test (Hui *et al.*, 2010), and instead, sites assumed spatially independent. This could lead to counterintuitive outcomes in the test (Hausdorf and Hennig, 2007). As a result, like Fuller and Enquist (2012), spatial autocorrelation of each species should be accounted for by the test.

Another essential thing that is ignored in the spatially implicit permutation null model test is the matching of environmental heterogeneity and species' habitat preference. There are possibilities that the patterns of species co-occurrences observed in ecological communities are due to environmental heterogeneity. That is, a species would prefer to be in a certain site or to coexist with another species in the same site because of the conditions of that environment. A species would prefer to live in a more favourable environment than the one it was occupying before. Therefore, its decision to live on that site would not necessarily be due to competition. During the hypothesis test, we might reject the null hypothesis and because environmental heterogeneity might also play a role in structuring ecological communities, it will be difficult for us to distinguish whether the effects of biotic interactions (e.g., competition) or environmental heterogeneity or both shape these communities.

Thus, to tease apart the role environmental heterogeneity from biotic interactions, the permutation test should also be allowed to reserve the association between species. This means sampling distribution should be generated by randomizing the observed pattern with the constraints on species association. That is, the species' habitat (or site) preference should be preserved. But since associated species would have similar preferences, species association should be preserved in the test. This project is thus to design a permutation null model test that can progressively include the spatial autocorrelation of species and the association between species. This would allow the role of environmental heterogeneity and aggregation in the structuring of ecological communities to be further investigated. A R package will be designed to implement both the spatially implicit null model tests of co-occurrence and newly designed approaches for the permutation test with the constraints on species autocorrelation and association. Real data will be used for model evaluation.

1.2 Specific Problem to be addressed

Statistical permutation tests or randomization tests are often used to generate sampling distributions by randomizing the observed data several times, where samples cannot be generated from the population. During such a process, certain information in the data can be altered. For instance, the spatial structure of the data can change. This can happen when two or more elements of the data which were far apart from each other are brought close together or vice versa. As a result, the level of their independence is subject to change. This change of data's level of independence or spatial autocorrelation during randomization tests to generate a sampling distribution can lead to counterintuitive results in the null model test. This is because the sampling distribution can be bias to have the null hypothesis rejected when it's suppose to be accepted, leading to commission of Type I error.

The main problem to be addressed therefore, is how to design a permutation null model test that can keep constant the spatial structure, and in particular, the spatial autocorrelation of species and the association between species during the randomization procedures to optimize on the accuracy of the statistical null model test of species co-occurrence.

1.3 Research Objectives

More precisely, there are two main objectives that will be addressed:

(i) Designing a permutation null model test that can progressively include the spatial autocorrelation of species and the association between species so that the role of aggregation and environmental heterogeneity can be further examined.

(ii) Developing a R package that can implement both classic null model tests of co-occurrence and newly designed approaches for the permutation test with the constraints on species spatial autocorrelation and association.

1.4 Limitations to the study

There is a limitation in regard to the data used in this study. In particular, the quantitative data describing the abundance of each species of birds from an empirical study which was done in Caribbean islands, are represented as binary data, with 1 indicating the presence of a particular species in a particular site and 0 its absence. This means, irrespective of the frequency of species occurring in a particular site, only a binary value (1) is used to represent them.

1.5 Research outline

This project is organised as follows: we review on the spatially implicit null model tests of species co-occurrence in chapter 2. Chapter 3 presents the newly designed approaches to null model testing. In particular, we discuss about the spatially explicit null model tests of species co-occurrence patterns, the procedures involved in carrying out these tests and their importance in regard to the accuracy of null models in community ecology. Chapter 4 presents a spatial null model package used to implement both the spatially implicit null model tests of co-occurrence and the newly designed approaches for the permutation test with the constraints on spatial autocorrelation and the association between species. Results of the comparison between the classic (spatially implicit) null models and the newly designed approaches (spatially explicit null models) are also presented in this chapter. We provide a conclusion in Chapter 5.

Chapter 2

Spatially implicit null models

2.1 Introduction

Null models were developed to test the effect of a mechanism or any cause of interest on, among other things, the structure of ecological communities, and to provide a reference point against which alternatives should be contrasted. There have been significant statistical tools used in the analysis of biogeographic and ecological data (Gotelli, 2001). Despite historical controversies surrounding their use (Gotelli, 2000), they have continued to unearth the solutions to many ecological problems which could not be puzzled out by the conventional statistical tests. In particular, some of the ecological problems which have been handled by null models include:

- Determining species abundance distributions. A species diversity null model have been used to address this problem, with the population processes determining species abundance clearly brought out (Gotelli *et al.*, 1996).
- Depicting the most effective resource−partitioning model from the different types proposed. The proposed resource−partitioning models include; *broken stick, dominance preemption, dominance decay, composite, geometric series, random fraction and random assortment* (Gotelli *et al.*, 1996).
- Test of nonrandomness (or independence) in the association between species within a community (Fuller and Enquist, 2012).
- Testing for the patterns of species negative co-occurrence, species clustering and nestedness in meta-communities (Hausdorf and Hennig, 2007).
- Determining the expected ratio of species to genus and other taxonomic ratios in a community where competition does not play a role in its structure (Gotelli, 2001), and

- Testing for the effects of inter-specific competition on the co-occurrence patterns of species in the community, among others.

The use of null models in solving this last ecological problem is of interest in this project, and it will be explored further in this chapter.

In summary, this chapter endeavours to provide a detailed definition of a null model, review on the spatially implicit null model tests, their history and use in depicting the forces behind the structure of ecological communities; how species-by-site matrices are used in the null model analysis of the patterns of species co-occurrences and the way in which the co-occurrence indices are used to summarized them. We will then discuss different spatially implicit null model algorithms used in randomizing the observed data to generate a sampling distribution and finalise with a step-by-step description of how a null model is used to test an hypothesis.

2.2 Definition of a null model

A number of expressions have been used to describe the definition and the use of a null model. A more detailed definition has been given by Gotelli *et al.* (1996). It states:

"*A null model is a pattern-generating model that is based on randomization of ecological data or random sampling from a known or imagined distribution. The null model is designed with respect to some ecological or evolutionary process of interest. Certain elements of the data are held constant, and others are allowed to vary stochastically to create new assemblage patterns. The randomization is designed to produce a pattern that would be expected in the absence of a particular ecological mechanism*"

2.3 History of null models

The use of null models to test the ecological theory's predictions has a long history. Their use came out of 1970's debates revolving around the role of inter-specific competition in structuring ecological communities (Diamond, 1975; Connor and Simberloff, 1979; Diamond and Gilpin, 1982; Gilpin and Diamond, 1982; Connor and Simberloff, 1983). Diamond (1975) argued that ecological communities are shaped and structured by the effects of inter-specific competition. He did his experiment using island assemblages and observed recurrent patterns that reflect inter-specific competition. Diamond (1975) designated these patterns as "assembly rules". Two of these rules were:

1. Certain species pairs (P, Q) may never coexist, so that replicate assemblages will harbour species P or species Q, but not both. These "checkerboard pairs" are a result of interspecific competition allowing one species to exclude another species that arrives later.

2. Due to inter-specific interactions, some species combinations (out of the $2^S - 1$ possible combinations formed from a total of S species) will be missing in nature since they are "forbidden combinations" that do not persist.

These interpretations were challenged by Connor and Simberloff (1979) by investigating the co-occurrence patterns expected in the absence of inter-specific competition, making use of the null models. As a result, a debate ensued that has lasted for over 3 decades (Sanderson *et al.*, 2009; Connor *et al.*, 2013). Since then, the use of null models to test ecological hypotheses has been valuable tools applied in many research areas including Veech (2012); Ackerly *et al.* (2006); Cornwell *et al.* (2006); Zimmermann *et al.* (2009); Bascompte and Melián (2005); Burns and Zotz (2010); Blüthgen *et al.* (2008); Adams (2007); Helmus *et al.* (2007); Ingram and Shurin (2009); Mouillot *et al.* (2008); Kembel (2009), among others.

2.4 Species co-occurrence

One of the very important uses of the null models have been analysing the species co-occurrence patterns (Gotelli, 2000). To do this, the empirical data is first summarised on the existence of a group of species on a batch of sites using the species-by-site matrices (Gotelli *et al.*, 1996). These matrices forms a fundamental element of analysing biogeographic and community ecological data (McCoy and Heck, 1987). We discuss how the data are organised in these matrices in the following section.

2.5 Species-by-site matrices

- The data are organized as a species-by-site matrix with R rows indexed by i and C columns indexed by j.
- Each row is a species and each column is a site.
- Entry b_{ij} in the matrix represents the occurrence (1) or absence (0) of species i in site j.
- Let us denote:

    i) O_i the total occurrences of species i across the sites (row total),

ii) S_j the total number of species occurring in site j (column total) and

iii) T the sum total of all occurrences of species in the matrix,

It follows that:

i) $O_i = \sum_{j=1}^{C} b_{ij}$

ii) $S_j = \sum_{i=1}^{R} b_{ij}$

iii) $T = \sum_{i=1}^{R} O_i = \sum_{j=1}^{C} S_j = \sum_{i=1}^{R} \sum_{j=1}^{C} b_{ij}$

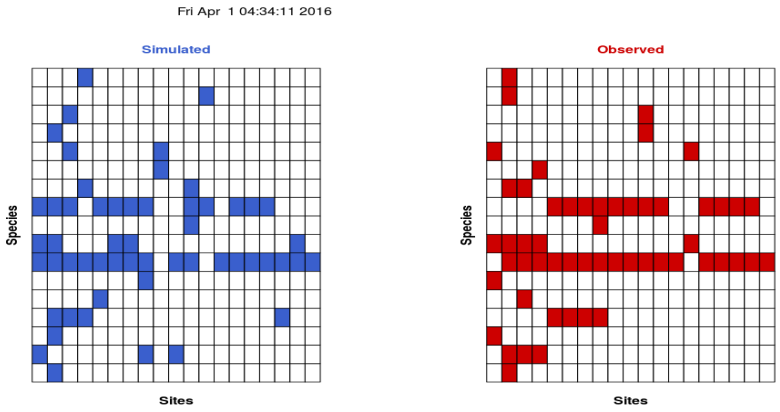


Figure 2.1: Example of a simulated species-by-site matrix (in blue) generated by randomizing the observed matrix (in red, like the data finches matrix in Table 2.1) with the constraints on row (O_i) and column (S_j) totals respectively (Nicholas J. Gotelli, 2016).

2.6 Co-occurrence Indices

Null model analysis requires that the data in the species-by-site matrices be summarised by a single number used in analysing patterns of species co-occurrence. Here, we present a detailed description of four co-occurrence indices, ranging from the early use of species combinations to the latest checkerboard score proposed by Stone and Roberts (1990).

2.6.1 Checkerboard Score

The term "checkerboard" was first used in community ecology by Diamond (1975) to describe the bird species that never coexisted in the Bismack Archipelego. It was argued that the checkerboard distributions was a consequence of species competition (Diamond, 1975), a surmise that the critics disputed saying the same patterns of species distribution could be generated by the null models

Table 2.1: A binary species-by-site matrix for West Indies finches

	Species	Cuba	Hispaniola	Jamaica	Puerto_Rico	Guadeloupe
1	*Carduelis_dominicensis*	0	1	0	0	0
2	*Loxia_leucoptera*	0	1	0	0	0
3	*Volatinia_jacarina*	0	0	0	0	0
4	*Sporophila_nigricollis*	0	0	0	0	0
5	*Melopyrrha_nigra*	1	0	0	0	0
6	*Loxigilla_portoricensis*	0	0	0	1	0
7	*Loxigilla_violacea*	0	1	1	0	0
8	*Loxigilla_noxis*	0	0	0	0	1
9	*Melanospiza_richardsoni*	0	0	0	0	0
10	*Tiara_olivacea*	1	1	1	1	0
11	*Tiara_bicolor*	0	1	1	1	1
12	*Tiara_canora*	1	0	0	0	0
13	*Loxipasser_anoxanthus*	0	0	1	0	0
14	*Saltator_albicollis*	0	0	0	0	1
15	*Torreornis_inexpectata*	1	0	0	0	0
16	*Ammodramus_savannarum*	0	1	1	1	0
17	*Zonotrichia_capensis*	0	1	0	0	0

Notes: The rows represents different finch species and the columns shows 5 of the 19 West Indies' major islands. Each cell of the matrix represents species' occurrence (1) or non-occurrence (0) on an island. Data from Gotelli and Abele (1982)

and consequently could not be a result of species interactions (Connor and Simberloff, 1979). This lead to a number of debates (Gotelli *et al.*, 1996) to ascertain whether the distribution of species would be different if there were no species interaction (Stone and Roberts, 1990). In 1990, the notion of species' checkerboard distribution was further expanded to a statistic (checkerboard score) used in ascertaining whether the spread of species over a collection of biomes is random (Stone and Roberts, 1990).

Following Stone and Roberts (1990)'s methods to calculate this statistic, we started with two species to see if their distribution was dependent on their interaction or random and make use of the species-by-site matrices with two species (rows) and m sites (columns). If CS_{ij} is the checkerboard score (C-score) for the two species in the m sites, then their C-score is expressed mathematically as:

$$CS_{ij} = (q_i - S_{ij})(q_j - S_{ij}) \tag{2.6.1}$$

where

q_i is the total occurrences of the first species,

q_j is the total occurrences of the second species and

S_{ij} is the total number of sites which harboured both species.

This represents a single checkerboard unit for a pair of species. To obtain the C-score for the whole pattern of species colonisation, we computed the average value of the checkerboard units per pair of species in the community. That is, if there are T different species in the community, the species pairs formed is given by $P = T(T-1)/2$, and therefore the C-score is expressed as:

$$C = \frac{1}{P}\sum_{j=0}^{T}\sum_{i<j} CS_{ij} \tag{2.6.2}$$

The C-score determines if the distribution of species across different sites is random. It is used as a metric which determines if biotic interactions played a crucial role in how species are spread across a collection of biomes. C-score is commonly used alongside some simulation algorithms to tell if biotic interactions can be inferred from different patterns of species-by-site matrices.

2.6.2 Number of Checkerboard Species pairs

The concept of "checkerboard distributions" has been used to describe every pair of species that never coexisted in the Bismack Archipelego (Diamond,

1975). Diamond (1975) argued that species who compete for the limited resources tend to keep away from each other, thereby creating a checkerboard pattern in the species-by-site matrices. During the null model analysis of the co-occurring patterns of species, the perfect checkerboard species pairs are counted and such a number is used to summarise the data in the species-by-site matrices. For a community in which species compete for resources, such checkerboard species pairs are significantly more than those anticipated by random chance (Diamond, 1975).

2.6.3 Number of Species combinations

Number of species combinations is another metric or co-occurrence index used in summarising the data in the species-by-site matrix before analysis. This metric is calculated by obtaining the total number of unique species combinations from different sites. A community with m species has 2^m species combinations with a combination of no species included (Pielou and Pielou, 1968). Since the total possible number of species combinations (2^m) is always greater than the total number of sites in most real matrices, this sets an upper limit on the total species combinations on both the observed and randomised matrices (Kesey-Bear, 2016). This index therefore only applies when you have lots of sampled sites.

To make an inference regarding the structure of ecological communities using this metric, if more species combinations exist than those expected by chance, this indicates a community or an assemblage which is not structured by inter-specific competition (Kesey-Bear, 2016).

2.6.4 Variance Ratio

Variance ratio (V.R) is the ratio of the variance of the column sums to the sum of the row variances (Gotelli, 2000), i.e.,

$$V.R = \frac{\text{Var(column sums)}}{\sum \text{(row variances)}}.$$

It measures the average covariance in association between all possible species pairs (Schluter, 1984). It was first used by Robson (1972) and later recommended by Schluter (1984) as a species co-occurrence index. To summarise the species-by-site matrix using this metric, the variances of both the column sums and row sums are first computed. Then the ratio of the former to the latter is obtained. For the equiprobably distributed sites and species which are independent and identically distributed, the expected variance ratio is equal to 1.0. A variance ratio with a value less than 1.0 indicates a strong negative

covariance between the pairs of species. Otherwise a value greater than 1.0 indicates a strong positive covariance between the pairs of species. The patterns of species co-occurrences does not determine the value of this metric unlike the previous co-occurrence metrics. Instead, the value of the variance ratio is determined by the matrix marginal totals (Gotelli, 2000). That means if the marginal totals of the observed matrix are maintained, sim9 algorithm (which will be discussed in the following section) cannot work.

In summary, although each of the four indices considered measure a slightly different aspect of species co-occurrence (Gotelli, 2000), C-score has been observed to have good statistical properties and is not prone to type I error, especially when used with sim9 or sim2 algorithms (Gotelli, 2000). It is therefore recommended. We will use this metric only in Chapter 3 to summarise the species-by-site matrix to be used in the analysis with the spatially explicit null models.

2.7 Randomization Algorithms

To carry out a permutation test, different algorithms have been used to generate a sampling distribution. The algorithms range from those that are highly constraint to those that impose almost no constraint at all (Gotelli, 2000). Of all the nine algorithms, some can lead to falsely rejecting the null hypothesis when it should actually be accepted (Type I error), while others are less prone to type I errors (like sim9). To guard against Type I errors, fixing row sums has been made a general rule during randomization of the observed data to generate a sampling distribution (Gotelli, 2000). It therefore becomes feasible to get accurate results if only four of these algorithms are used to generate the 'random' matrices (Gotelli, 2000). One of the four algorithms is sim9 with the histogram of the C-scores of the simulated data shown in Figure 2.2. We present a detailed description of these algorithms in the following section.

sim1

Equiprobable rows, equiprobable columns

This algorithm was first used by Sokal and Rohlf (1995). It can be thought of as the most 'random/null' relative to other randomization algorithms since all the species and sites are assumed to be equiprobable. Sokal and Rohlf (1995) used it during the randomization tests where all the data combinations are equally likely. That is, the probability $P(b_{ij})$ of cell occupancy is given by

$$P(b_{ij}) = \frac{1}{R_T C_T},$$

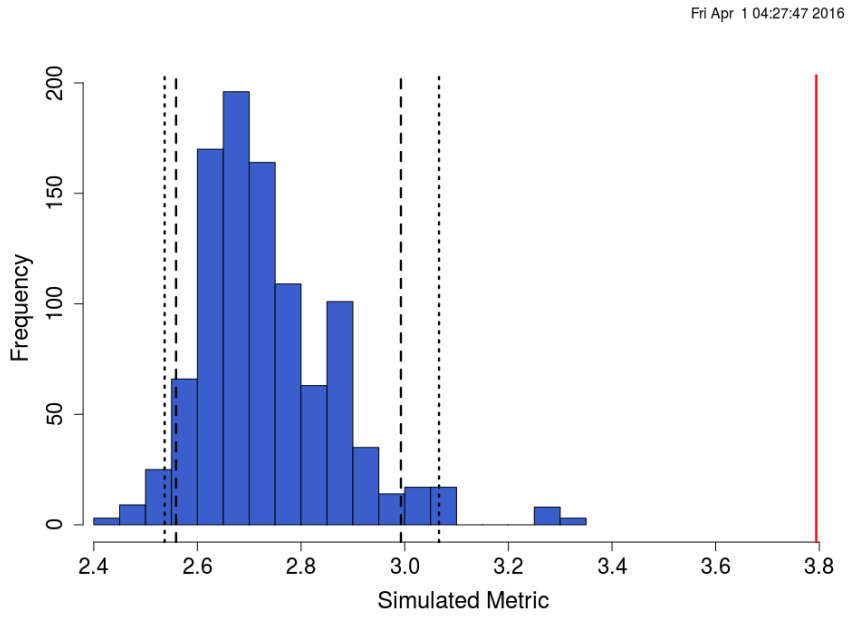


Figure 2.2: Histogram of Simulated metrics. The C-score of the observed data is indicated by the red vertical line. The cut-points for the 95% two-tailed test are indicated by the short-dash vertical lines. The long-dash vertical lines indicate the cut-points for the 95% one-tailed test.

where

R_T is the total number of rows in the matrix,

C_T is the total number of columns in the matrix and

b_{ij} is an element (or cell occupancy) in the i^{th} row and j^{th} column.

Irrespective of the co-occurrence index used along with this algorithm, the results of the tests are highly unreliable due to high error rates. As such, it should not be used (Gotelli, 2000).

sim2

Fixed rows, equiprobable columns

This algorithm relies on the total number of sites (columns). It makes an assumption that the sites are equally probable. According to Gotelli (2000), it is similar to a simple model of an assemblage where species occupy sites 'randomly'. However, it is constraint by fixing the totals of species occurrences. Every cell occupancies are governed by the probability $P(b_{ij})$ given by

$$P(b_{ij}) = \frac{1}{C_T},$$

where b_{ij} and C_T are as defined in sim1 algorithm. When used with any of the 4 co-occurrence indices, the results were satisfactory (Gotelli, 2000), and is therefore recommended.

sim3

Equiprobable rows, fixed columns

This Algorithm is the inverse of sim2 algorithm. The total number of species in every site is fixed. Unlike sim1 where all the species and sites are equally probable, this algorithm randomizes only all the species equiprobably. The probability of every cell occupancy is given by

$$P(b_{ij}) = \frac{1}{R_T},$$

where b_{ij} and R_T are as defined in sim1 algorithm. Like sim1, it is highly prone to statistical errors when used with all the indices (Gotelli, 2000). Due to this, it is not recommended.

sim4

Fixed rows, proportional columns

sim4 algorithm randomizes the species-by-site matrix with the constraint on the total occurrences of species observed. It fixes the total occurrences of every species while randomizing the species occurrences among the sites. Unlike sim2, sites are not equiprobable, and every cell has the probability proportional to the column totals observed. Therefore, probability is given by

$$P(b_{ij}) = \frac{T_j}{N}$$

where

T_j is the sum total for j^{th} column,

N is the sum of all the occurrences in the matrix and

b_{ij} is as defined in sim1 algorithm.

This algorithm has been equated to a "random placement" model of species on sites (Coleman *et al.*, 1982). It has been proved to work well with two of the co-occurrence indices (i.e., the Variance-ratio and the number of species combinations). However, it is prone to statistical errors when used with "Checkerboard score" (C-score) and the "number of Checkerboard Species pairs" (Gotelli, 2000). As such, it should only be used with the former two indices.

sim5

Proportional rows, fixed columns

This algorithm simulates the species-by-site matrix with the constraint on the column totals, i.e., species richness per site is fixed (Gotelli, 2000). It is the inverse of sim4 algorithm. The occurrence probability $P(b_{ij})$ of the species is proportional to the observed frequencies of the species occurrence (Gotelli, 2000), and is given by

$$P(b_{ij}) = \frac{S_i}{N},$$

where

S_i is the sum total for the i^{th} row and

N is as defined in sim4 algorithm.

When used with any of the four co-occurrence indices, it portrays high error rates. It is therefore not recommended.

sim6

Equiprobable rows, proportional columns

This algorithm simulates the species-by-site matrix with the assumption that all the species are equiprobable and the site probabilities of occurrence are proportional to the species richness observed per site (Gotelli, 2000). In other words, site probabilities are proportion to the sum total of every column. The occurrence probability $P(b_{ij})$ is given by (Gotelli, 2000)

$$P(b_{ij}) = \frac{T_j}{N \times R_T}$$

where T_j, N and R_T are as defined in sim4, sim5 and sim1 algorithms respectively. There is high error rates when used with the variance ratio and the number of species combinations indices. The errors rates are also not satisfactory when used with the number of checkerboard species pairs and the checkerboard score. It is therefore not recommended.

sim7

Proportional rows, equiprobable columns

This algorithm assumes all the sites to be equiprobable with the species differing in occurrence, i.e., there is a variation in occurrence probabilities for

different species. This variation is in proportion to the sum totals of every row (Gotelli, 2000). The occurrence probabilities $P(b_{ij})$ are given by

$$P(b_{ij}) = \frac{S_i}{N \times C_T},$$

where S_i and N are as defined in sim5 and C_T is as defined in sim1. It performs poorly like sim6 algorithm, and is therefore not recommended (Gotelli, 2000).

sim8

Proportional rows, proportional columns

In this algorithm, sites and species differ in suitability (Gotelli, 2000). It assumes that neither the species nor the sites are equiprobable. The occurrence probabilities $P(b_{ij})$ are conditional on the marginal totals of both the species and the sites (Gotelli, 2000). This probability is given by

$$P(b_{ij}) = \frac{S_i \times T_j}{N^2},$$

where S_i and N are as defined in sim5, and T_j is as defined in sim4. It performs well with the variance ratio and the number of species combinations indices. However, it is prone to error when used with the number of checkerboard species pairs and the checkerboard score. It should therefore only be used with the former two indices.

sim9

Fixed rows, fixed columns

This algorithm is highly recommended (Gotelli, 2000). It maintains row and column totals, implying degenerate matrices are not produced. It provides a modified version of the Connor and Simberloff (1979)'s algorithm, which received a lot of criticisms. Unlike the other algorithms, it can be used to detect patterns in noisy data when used with the checkerboard score. It cannot be used however with the variance ratio since it is determined exclusively by column and row totals. The occurrence probabilities $P(b_{ij})$ are not applicable in this case. This is because sim9 cannot be simulated by filling an empty matrix (Gotelli, 2000).

2.8 Null model as a test of hypothesis

We illustrate how a null model is used to test an hypothesis in this section. In layman's terms, an hypothesis is a statement which can be true or false and is subject to approval or testing to derive a conclusion. It is a proposition or

a theory set forth to explain the occurrence of an observed event or a phenomenon. Hypothesis testing is commonly used to make decisions in statistics based on data or observations, i.e., based on statistics, hypothesis testing is equivalent to trying to establish if an observed phenomenon is likely to have happened.

Most null models have been used as statistical tests of null hypotheses. In some instances, some have been used in simulating an ecological process with no significance test accompanied. To ascertain how a null hypothesis can be tested using a null model, let us consider the following.

1. Null hypothesis testing relies on a sampling distribution from where the test statistic to be compared with the observed parameter is calculated. This probability distribution presents the distribution of the frequencies of a range of different outcomes that can occur for the parameter of the population.

2. Normally, samples are taken independently from a known population.

3. If that is not possible, a model is used to randomize the observed data several times to generate a sampling distribution.

4. The samples generated in such a manner forms a probability distribution called a null distribution. From here, two approaches can be used to obtain the p value. Either a direct approach where the p value is computed as a fraction of the number of more extreme values than the value observed, and the total number of randomizations/permutations performed, or an indirect approach where the null distribution is approximated by a known distribution and significance test performed to compare the observed value to the values in the null distribution. The p value is obtained from the test of significance in this case.

Figure 2.3 presents a step-by-step procedure involved testing a null hypothesis using a null model. Whereas a null model formalizes a particular null hypothesis, the two terms are sometimes viewed as synonymous. However, a few distinctions between them are presented in Figure 2.4.

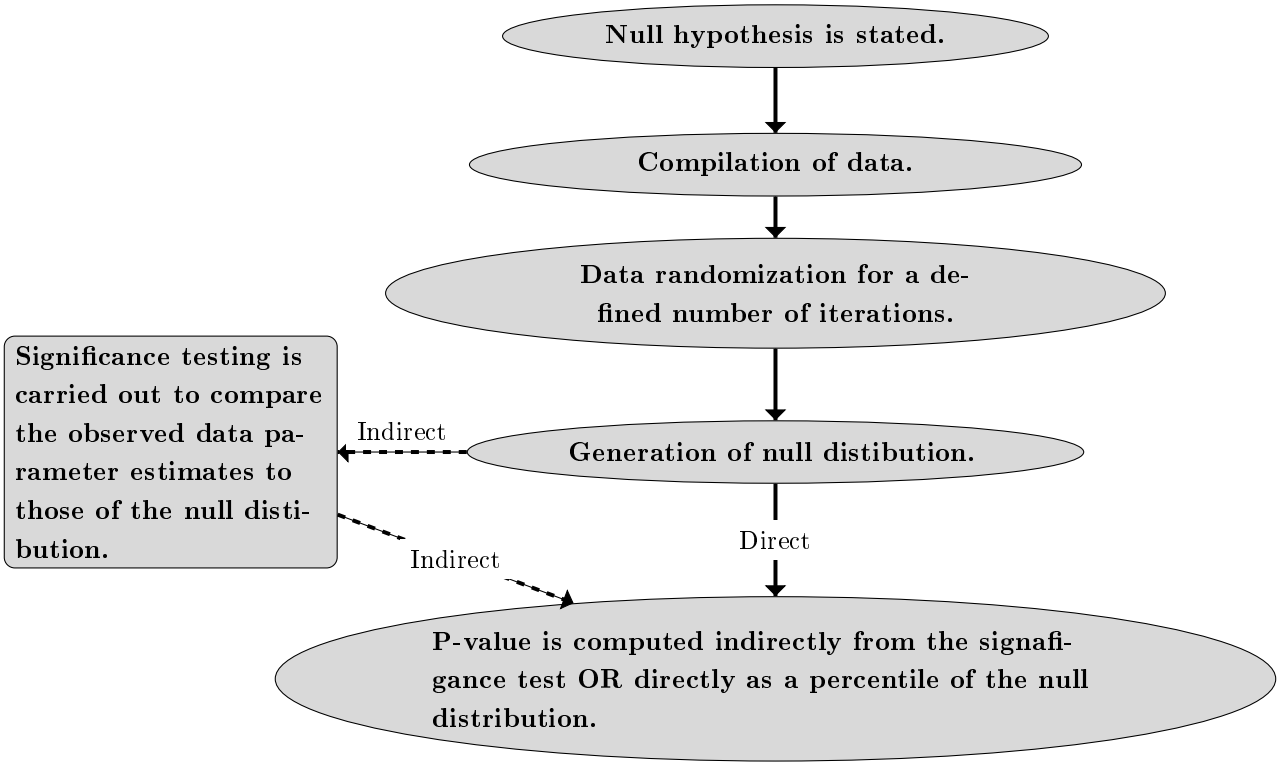


Figure 2.3: Procedures involved in testing a null hypothesis using a null model. The difference between a direct and indirect tests lies in the computation of a p value. There is an extra step of using statistical inferential test to compute a p value in indirect test, unlike the direct test where the p value is directly obtained from the distribution of null values. *Source:* (Veech, 2012)

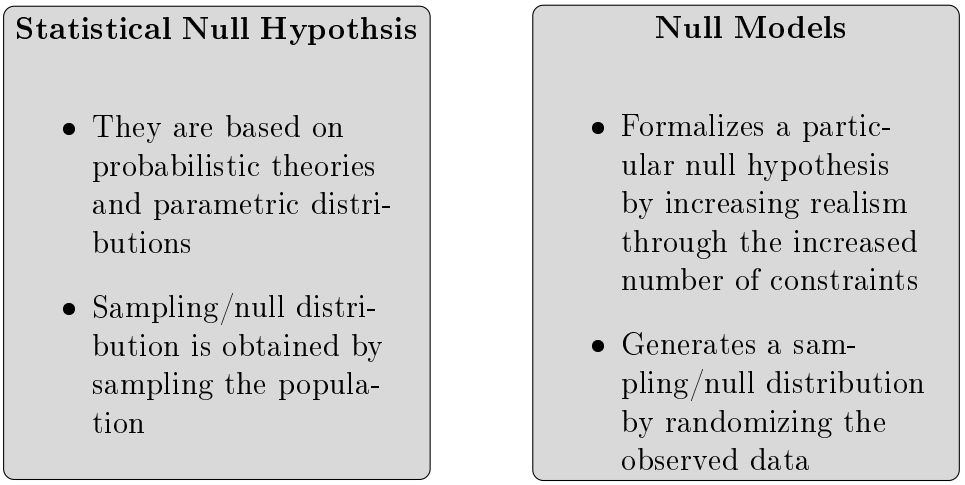


Figure 2.4: While the sampling distribution in the null hypothesis comes from a known distribution, null models rely on randomization techniques to generate a sampling distribution (MBASKOOL, 2016).

2.9 Summary

We have presented the classic (or spatially implicit) null model procedures in this chapter. In particular, various null model algorithms and metrics (or co-occurrence indices) used to summarise the species-by-site matrices using a single number have been reviewed. Figure 2.5 gives a summary of the steps involved in null model testing generally.

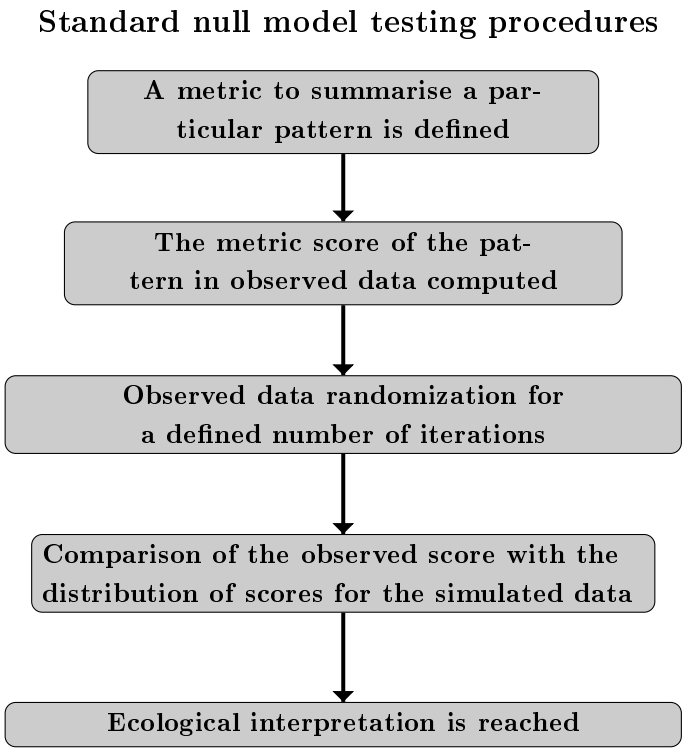


Figure 2.5: A flow chart summarising the standard null model testing procedures (Gotelli and Ulrich, 2012).

Chapter 3

Spatially explicit null models

3.1 Introduction

All the null model procedures described so far do not account for the spatial autocorrelation of species distributions and the association between species. This could lead to counter-intuitive results in the null model test. The latter problem will be dealt with in this chapter.

Null hypothesis tests, as mentioned in the last section, relies on observations being independent from one another during permutation tests to generate a sampling distribution. For any process, the null hypothesis will either be true and therefore acceptable or false and therefore rejectable. If hypothesis is rejected when it should actually be accepted, an error (called type I error) is committed. One of the major reasons one can commit this type of error is dependent sampling. That is, generating samples from the population or by re-sampling the observed data in a biased non-random manner, such that generating any sample data at any point in time depends on the previous data sampled.

If the process generates unbiased and random samples (like permutation tests), certain characteristics of the data (like independence) should not be affected by the same process, especially if the hypothesis test involved is a test of independence. Otherwise the results of the test would still be compromised. Therefore, for any hypothesis test to be presumed valid, all of its assumptions, including the independent non-biased sampling procedures, have to be observed, failure to which the inferences made from the results of the test would be entirely wrong and therefore unreliable.

In addition, to test for independence, certain constraints have to be satisfied if the sampling procedures employed to generate the sampling distribution interferes with the spatial structure or the level of independence of the observed

data. For instance, if we want to test the effect of inter-specific competition on the structure of ecological communities, the hypothesis to be tested would be:

> $\mathbf{H}_0$: Species are independent in their decision to coexist. Patterns of species co-occurrences are therefore formed by random chance. That is, inter-specific competition does not play any role in the species co-occurrence patterns. Since the patterns are random, the C-score of the observed matrix is no different from the C-scores of the sample matrices (generated by the permutation test).
>
> $\mathbf{H}_1$: Species are not independent in their decision to coexist with other species in the same site (or community). Their patterns of co-occurrences are therefore not formed by random chance but due to inter-specific competition. That is, if two or more different species compete for the limited resources, they try to avoid each other as much as possible. If there is no competition between them, they will coexist peacefully with one another. Since the patterns change due to competition (and not by random chance), the C-score of the observed matrix is totally different from the C-scores of the sample matrices. By randomizing the observed matrix to generate a sampling distribution, the samples generated will have a different structure from the one observed and their C-scores will consequently be different from the C-score of the observed matrix.

For this hypothesis to be tested correctly, the randomization procedures (or permutation test) used to generate a sampling distribution should keep constant the level of independence of species (since the test involved is that of independence). That is, a constraint should be imposed such that the level of independence of species in the observed data is the same as the level of independence of species in the sample data. In this way, the results of the test will lead to a valid inference regarding the species co-occurrence patterns.

An example of keeping a certain characteristic in the data constant to generate a valid inference regarding the hypothesis in question would be to keep the temperature of the stove constant in testing the hypothesis that water's temperature reaches a $100°C$ after 30 minutes of heating using a stove emitting heat at $300°C$. If the same water is heated using a stove emitting a different amount of heat other than $300°C$, we will be shift in rejecting the null hypothesis (and therefore committing type I error) when our water temperature shows something different from $100°C$ after 30 minutes of heating. One would say we cannot make an inference regarding our hypothesis by using just one sample. But still, if the same procedure is done 1000 times, each time altering the stove's temperature, the sampling distribution of 1000 water temperatures would lead to falsely rejecting the null hypothesis. This is because the time it takes the water to boil depends entirely on the amount of heat emitted by the

stove, assuming that the same quality of the material is used to heat the water. In this case, the temperature of the stove should be kept constant both in the hypothesised experiment (i.e., under the null hypothesis) and in the sample experiment. This example brings out clearly the importance of certain characteristics about the data that should be held constant during the permutation test.

However, classic null model tests (herein referred to as spatially implicit null model tests), as seen in the previous chapter, do not take into consideration the effect of permutation tests in altering the spatial structure of the resulting samples (from the observed data). This interferes with the level of independence of species in the sample data. This means after the permutation tests, the level of independence of the sample data will not be the same as that of the observed data, and therefore using the results of the test which relies on the sampling distribution to make inferences about the hypothesis in question, would lead to counter-intuitive outcomes. Thus, one has to be keen during sampling to ensure that all the assumptions of the test, including independence, are not violated. Consequently, these features of the data which affects the outcomes of the test (if altered) should be kept constant during the permutation test so that the outcomes of the test can be fully attributed to the hypothesis in question.

To keep these features constant during the permutation tests, one has to quantify them and impose the constraints that will keep the same quantities constant during the randomization procedures. Since the level of species independence in their decision to colonise sites affects the structure of ecological communities, we will concentrate on keeping the level of species independence constant during the permutation test to generate a sampling distribution. The level of species independence can be best described and quantified using a concept called spatial autocorrelation. This concept can be measured and quantified using Moran's I coefficient. Since the analysis of ecological and biogeographical data are mostly affected by spatial autocorrelation, this concept has been a source of interest for many researchers, for example Fuller and Enquist (2012); Hausdorf and Hennig (2007); Diniz-Filho *et al.* (2003); Lichstein *et al.* (2002); F Dormann *et al.* (2007); Griffith and Chun (2015); Mathiba and Awuah-Offei (2015); Liu *et al.* (2015); Westerholt *et al.* (2015); Melecky (2015), among others. We discuss this concept and the metric of its measurement in greater detail in the following section.

3.2 Spatial Autocorrelation

Spatial autocorrelation is a phenomenon in which the similarity of two variables X and Y diminishes as the distance between them increases. This explains

why observations made at nearby locations may be dependent on each other, as opposed to observations made at locations farther apart. For instance, measurements made at farther locations may be more distant in value than the measurements made at locations nearby. This phenomenon is well illustrated in Figure 3.1.

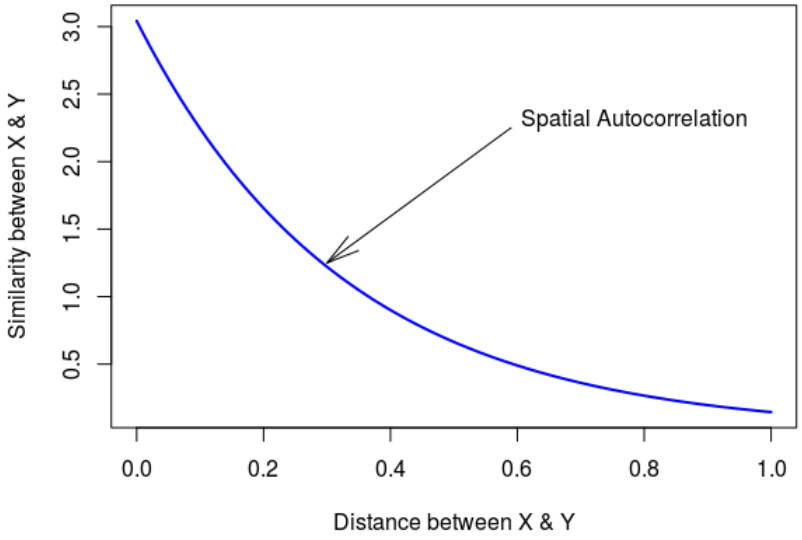
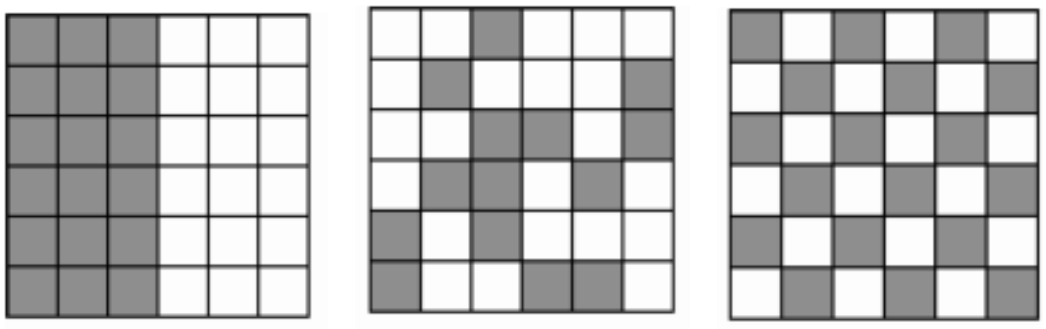


Figure 3.1: Distance decay of similarity illustrating spatial autocorrelation.

Figure 3.1 echoes Tobler (1970)'s "first law of geography", which states that "everything is related to everything else, but near things are more related than distant things". Spatial autocorrelation can either be positive or negative.

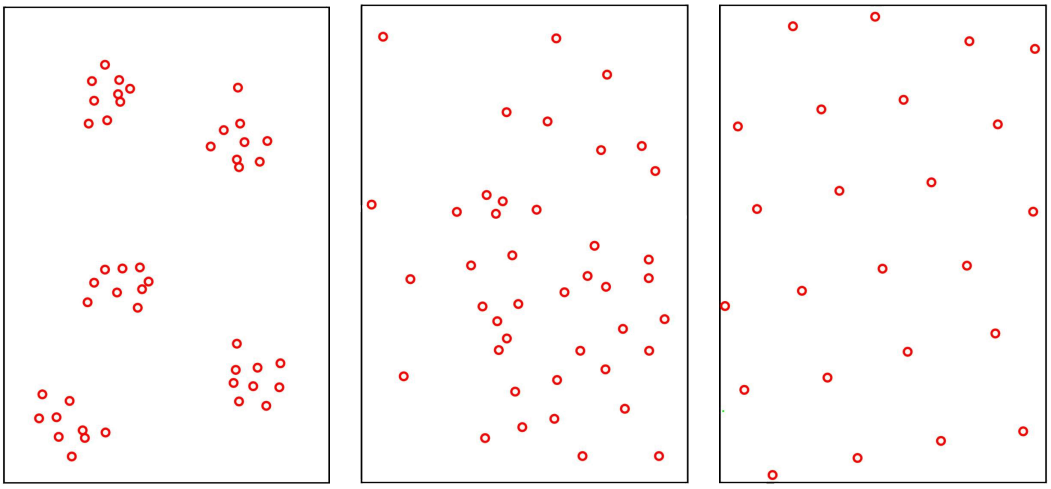
- Positive spatial autocorrelation (Figure 3.3a) is said to occur when similar values are observed to cluster together in a map. In such a case, the pattern of the relationship is said to be clustered.
- Negative spatial autocorrelation (Figure 3.3c) is said to occur when dissimilar values cluster together in a map. If there is negative spatial autocorrelation in the data, the pattern of the relationship is said to be dispersed.
- If there is no spatial autocorrelation (Figure 3.3b), the pattern will appear to be random. That is, the pattern is neither clustered nor dispersed.

These patterns are illustrated by Figures 3.2, 3.3 and 3.4. To quantify and describe the nature of their relationship, we will consider the Moran's I coefficient as the measure of spatial autocorrelation in the following section.



(a) Positive spatial autocorrelation (b) No spatial autocorrelation (c) Negative spatial autocorrelation

Figure 3.2: Spatial autocorrelation



(a) Clustered (b) Random (c) Dispersed

Figure 3.3: A visualization of the spatial patterns of species with different types of spatial autocorrelation. *Source*: (Wikipedia, 2016)

Figure 3.4: Types of spatial distribution for polygon data.

(a) Clustered (b) Random (c) Dispersed

Notes: (a) illustrates a strong positive spatial dependency. There is no spatial dependency in (b). A strong negative spatial dependency is evident in (c). *Source*: (GITTA, 2016)

3.2.1 Moran's I coefficient

Moran's I coefficient is the most common metric used in measuring spatial autocorrelation. It is expressed mathematically as (Moran, 1950):

$$I = n\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\sum_{i=1}^{n}(z_i - \bar{z})^2}, \tag{3.2.1}$$

where

n is the total number of observations,

z_i is the observation at location i,

z_j is the observation at location j,

$\bar{z}$ is the average of all the observations, and

w_{ij} is the distance weights matrix. This value indexes the location of i relative to j.

The values of this coefficient range from -1 to 1, where -1 is an indication of perfect dispersion and 1 is an indication of perfect correlation. A random spatial pattern is indicated by a zero value.

The spatial autocorrelation measurement using this tool is based simultaneously on characteristic locations and feature values.

3.2.2 Importance of spatial autocorrelation

Spatial autocorrelation has been a key feature in spatial analysis and ecological modelling for many researchers in ecology and biogeography, for example, Betts *et al.* (2006); Bahn *et al.* (2006); Zierahn (2012). The point of concern has been to account for the spatial interdependencies in ecological and biogeographic data. For instance, statistical tests which assume independence of the observations can lead to counter-intuitive inference if the independence assumption is violated. This violation can be as a result of failing to account for the spatial autocorrelation in the models.

Spatial autocorrelation has also been important in the analysis of communities or clusters and the dispersion of disease and ecology. The disease can be seen as an isolated case or spreading with dispersion with the help of the spatial autocorrelation analysis (GIS, 2016).

To make valid inferences from the results of the null model test on the co-occurrence patterns of species, spatial autocorrelation will be fixed during the randomization of the observed data to generate a sampling distribution. In particular, a small number ϵ will be defined such that

$$\left| \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \sum_{i=1}^{n} (z_i - \bar{z})^2} - \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(\hat{z}_i - \bar{\hat{z}})(\hat{z}_j - \bar{\hat{z}})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \sum_{i=1}^{n} (\hat{z}_i - \bar{\hat{z}})^2} \right| \leq \epsilon \quad (3.2.2)$$

i.e.,

$|I - \hat{I}| \leq \epsilon$, where

I is the Moran's I value of the observed data,

$\hat{I}$ is the Moran's I value of the simulated data,

$\hat{z}_i$ is the observation at location i after simulation,

$\hat{z}_j$ is the observation at location j after simulation and

$\bar{\hat{z}}$ is the mean value of the observations after simulation.

The whole null model process of the permutation test with the constraint on the spatial autocorrelation will be discussed in section 3.5.1.

3.3 Species Association

The association between two species can influence their decision to either coexist in the same site or live separately in different sites. This means the structure of ecological communities is influenced by the ecological association between species. To test the effect of inter-specific competition in structuring ecological communities using null models, species association therefore have to be kept constant during the permutation test.

We consider the measurement of this feature using a metric called association index proposed by Dice (1945), in the following section.

3.3.1 Association Index

To quantify the amount of ecological association between different species, association index will be used. For the two species i and species j, the amount of ecological association between them is given by

$$AI_{ij} = \frac{2S_{ij}}{r_i + r_j}, \tag{3.3.1}$$

where

r_i is the total occurrences of species i across all the sites,

r_j is the total occurrences of species j across all the sites and

S_{ij} is the total co-occurrences of both species i and j.

AI_{ij} represents the single association index or unit between the two species, i and j. To get the association index for the entire species colonisation pattern, we calculate the mean number of the association units for every pair of species in the assemblage. That is, all the association units are summed and the result divided by the total number of species pairs. This is expressed mathematically as

$$AI = \frac{1}{P} \sum_{j} \sum_{i<j} \frac{2S_{ij}}{r_i + r_j}, \tag{3.3.2}$$

where

$P = \frac{M(M-1)}{2}$ is the total number of species pairs formed by a total of M species and

S_{ij}, r_i and r_j are as defined in equation 3.3.1

This index will be fixed as the constraint during the randomization of the observed data to generate a sampling distribution. In particular, a small number ϵ will be defined such that

$$\left| \frac{1}{P} \sum_{j=1}^{M} \sum_{i<j} \frac{2S_{ij}}{r_i + r_j} - \frac{1}{P} \sum_{j=1}^{M} \sum_{i<j} \frac{2\hat{S}_{ij}}{\hat{r}_i + \hat{r}_j} \right| \leq \epsilon \tag{3.3.3}$$

i.e.,

$|AI - \hat{AI}| \leq \epsilon$, where

AI is the association index of the observed data,

$\hat{AI}$ is the association index of the simulated data,

$\hat{r}_i$ is the total number of occurrences of species i across all the sites after simulation,

$\hat{r}_j$ is the total number of occurrences of species j across all the sites after simulation and

$\hat{S}_{ij}$ is the total number of co-occurrences of both species i and j after simulation.

The whole null model process of the permutation test with the constraint on ecological association between species will be discussed in section 3.5.2.

For the above constraints, ϵ represents an error rate of utmost 0.1%. This implies a similarity of atleast 99.9% between the observed values and the accepted simulated values used to generate the sampling distribution. The acceptance rate is thus atleast 99.9% and ϵ is therefore default in the R package to be presented in Chapter 4.

3.4 Null model procedures

Having described the two factors (spatial autocorrelation and species association) affecting the outcomes of the null model test in the previous sections, we now consider the steps involved in testing the hypothesis. The procedure to be used depends on the simulation algorithms used to generate a sampling distribution. We describe these procedures in detail in section 3.5. Generally, the steps involved are

- Design the algorithms that generate the null or the sampling distribution (i.e., the distribution of the C-score values of the simulated matrices) from the observed data without violating any of the assumptions of the test. In this case, the algorithm must keep constant both

  (i) the spatial autocorrelation of individual species and

  (ii) the amount of ecologic association between species

  using equations (3.2.1) and (3.3.2) during the randomization of the observed data to generate a sampling distribution, so that the role of aggregation and environmental heterogeneity can be further examined on the effect they have on the structure of ecological communities. This is because the outcomes of the test can only be attributed to inter-specific competition if all the other factors affecting the structure of ecological communities are held constant.

- Formulate the mathematical expression of the hypothesis (refer to section 3.6) and test it using the methods discussed in section 3.6.1).

- Finally, appropriate conclusions are drawn from the results of the test.

3.5 Simulation algorithms

Two algorithms will be considered in this section:

- Spatial algorithm using Moran's I coefficient as the permutation constraint

- Association algorithm using Association Index as the permutation constraint

We describe the steps involved in each below.

3.5.1 Spatial algorithm (Spatial1)

This algorithm generates a sampling distribution by fixing the spatial autocorrelation of individual species during the permutation test. The following are the steps involved.

i) Transpose the species-by-site matrix.

ii) Select all the columns of the transposed matrix above.

iii) For every column, compute the spatial autocorrelation using the Moran's I coefficient.

iv) Randomize every column many times and extract the random samples generated such that their Moran's I values are equivalent to the Moran's I values of the original columns.

v) From the random samples generated for a single column in (iv) above, select one sample randomly and replace the original column with it. Do this for all the other columns to form a new matrix.

vi) Return this new matrix and transpose it to have a new species-by-site matrix.

vii) Compute the C-Score of the output.

viii) Repeat all the steps above 1000 times to generate a sampling distribution of the C-score values. Simulating the data 1000 times increases the precision of the test. Accuracy and precision of the permutation test increases with the increasing sample size upto a certain point (mostly 1000), beyond which accuracy and precision of the test remains constant.

ix) Perform hypothesis testing using either an indirect approach if the sampling/null distribution follows a known probability distribution function, or direct approach (if the sampling distribution is unknown) to obtain the p value and the confidence intervals within which the null hypothesis should be accepted or rejected.

x) Interpret the p value and consequently carry out ecological interpretation and make appropriate conclusions regarding the species co-occurrence patterns.

3.5.2 Association algorithm (Spatial2)

This algorithm generates a sampling distribution by fixing the ecological association of species during the permutation test. It follows all the steps as outlined in subsection 3.5.1, except that association index is used in place of the Moran's I coefficient.

3.6 Mathematical expression of the hypothesis

As stated in section 3.1, the hypothesis test involved is that of independence. We present a mathematical expression of this hypothesis:

$$\begin{cases} H_0 : C_s = C_{obs} & \text{(patterns are random)} \\ H_1 : C_s \neq C_{obs} & \text{(patterns are dependent on inter-specific competition)} \end{cases}$$

where

C_{obs} is the C-score value of the observed matrix

C_s is the hypothesised value of the C-score

3.6.1 Testing of the hypothesis

There are two approaches in testing the above hypothesis. These are Direct and Indirect approach (refer to Figure 2.3).

Direct Approach

This approach does not rely on any known probability distribution like the normal or student t distribution. Instead, it computes the confidence interval (within which the null hypothesis will be accepted if the observed value falls under) by obtaining both the 2.5^{th} and 97.5^{th} percentiles of the sampling distribution assuming the significance level of 5%. If the C-score value of the observed matrix will fall outside this confidence interval, the null hypothesis will be rejected. The 2.5^{th} and 97.5^{th} percentile values are called critical values. If the histogram of the sampling distribution portrays a bell-shape, these regions of acceptance and rejection are well illustrated in Figure 3.5.

With the critical values, we are 95% confident that our decision to either reject or accept the null hypothesis is correct. To get the actual measure of how extreme the C-score of the observed data is relative to the sampling distribution, we compute the probability of obtaining an equal to or "more extreme" result than the C-score of the observed data, with the null hypothesis pre-assumed true. This probability is what is termed as the p value.

To compute the p value using a direct approach, the following expression is used.

$$p \text{ value} = \frac{\text{No. of } |C_{sim}| \text{ values} \geq |C_{obs}| \text{ value}}{n},$$

where

n is the total number of simulations

C_{obs} is the C-score value of the observed matrix and

C_{sim} values are the C-scores of the simulated matrices

Now if p value is less or equal to the defined level of significance, this provides enough evidence against the null hypothesis. Otherwise the null hypothesis

will be accepted. Similarly if the C-score of the observed data falls outside the confidence interval created by the critical values, we reject the null hypothesis, and vice versa.

Indirect Approach

Unlike the direct approach, this approach uses a known probability distribution to approximate the null/sampling distribution. For instance, depending on the number of simulations performed to generate the sampling distribution, a normal probability distribution can be used. Using the null model to generate a null/sampling distribution (refer to Figure 2.3), the following are the steps involved in computing both the confidence interval and the p value using an indirect approach, assuming a normal (Gaussian) distribution.

First, for the confidence interval, the critical values (C.V) are given by

$$C.V = C_s \pm Z_{\frac{\alpha}{2}} \delta_{\bar{C_{sim}}}, \tag{3.6.1}$$

where

C_s is the hypothesised C-score value. Under the null hypothesis, it is equal to the C-score of the observed matrix.

$\delta_{\bar{C_{sim}}}$ is the standard error of the C-score values forming the null distribution.

$Z_{\frac{\alpha}{2}}$ value is read from the standard normal tables.

Using the central limit theorem, if the number of random samples n is greater or equal to 30, it implies that

$$\bar{C_{sim}} \sim N(\bar{\bar{C_{sim}}}, \delta^2_{\bar{C_{sim}}}).$$

That is, sample means of the C-score values of the simulated matrices (assuming the sample mean is computed for every n simulations done) will be normally distributed with mean $\bar{\bar{C_{sim}}}$ and variance $\delta^2_{\bar{C_{sim}}}$.

Now to accept or reject the null hypothesis, the hypothesised C-score value is compared with the critical values computed. If the hypothesised value falls within the interval created by the critical values, the null hypothesis is accepted. Otherwise, it is rejected.

Alternatively, instead of using the critical values (which uses the hypothesised value of the C-score) defined in equation 3.6.1, the $\bar{C_{sim}}$ values can be

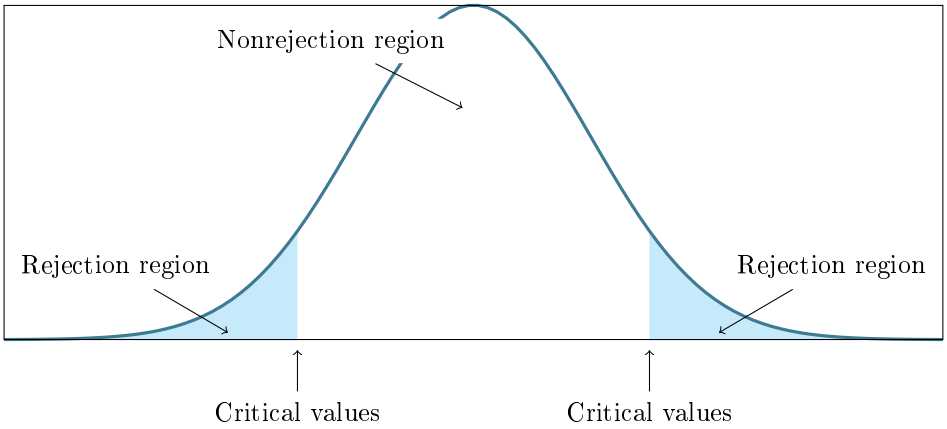


Figure 3.5: A visualization of a null distribution (assuming it is Gaussian/normal) and the regions within which the null hypothesis should be accepted or rejected depending on where the observed C-score value lies, using two-tailed test.

standardised using the following expression

$$z_c = \frac{C^{-}_{sim} - C^{=}_{sim}}{\delta_{C^{-}_{sim}}}; \qquad (c \text{ in } z_c \text{ signifies computed value}) \tag{3.6.2}$$

and $Z_{\frac{\delta}{2}}$ value obtained from the standard normal tables. Since the normal curve is symmetrical, $-Z_{\frac{\delta}{2}}$ and $Z_{\frac{\delta}{2}}$ forms the critical values within which the null hypothesis is accepted.

Lastly, to compute the p value using the indirect approach, the standard normal tables is first used to find out the percentage of the standard normal distribution that falls between the computed z value and positive infinity. This is equivalent to computing the probability that the Z variable is greater than the computed z value. That is $P(Z > z_c)$. Since our test is two-tailed, this value is multiplied by 2 to account for the lower tail. In sum, p value of a two-tailed test is given by

$$p \text{ value} = P(Z < -z_c \text{ or } Z > z_c)$$

The p value is interpreted the same way as in direct approach.

Which approach to use: Direct or Indirect?

If the null distribution is Gaussian/normal, the p values from both the direct and indirect tests are approximately the same. However, if the null distribution is non-Gaussian, using an indirect test increases the statistical error (Type I and Type II) rates, compared to direct test. Because of the assumption of normality which must be adhered to, it suffices to always use a direct test.

The form of the null distribution assumptions are not made in direct test to compute both the p value and the confidence intervals. Instead, the p value is simply the proportion of the samples which are "more extreme" than the value observed, and the confidence intervals are simply computed by taking some percentile of the sample distribution regardless of whether it follows a normal probability density function or not (Veech, 2012).

3.7 Summary

We have looked at the spatially explicit null models in this chapter. In particular, we have incorporated the spatial autocorrelation and species association in the statistical null model test of co-occurrence. Consequently, two null model algorithms have been designed;

- the spatial algorithm – that incorporates the spatial autocorrelation of species in the null model and
- the association algorithm – that accounts for the ecological association between species pairs in the ecological community.

The two algorithms can be thought of as two spatial null models. Unlike the classic (spatially implicit) null models, the spatial null models do not violate any of the assumptions of the hypothesis test. By fixing the spatial autocorrelation and species association during the permutation test to generate a null distribution, the test's assumption of independence has been observed. It therefore suffices to conclude that spatial null models are more reliable than their classic counterparts.

Chapter 4

Spatial null model package

4.1 Introduction

To implement both classical null model tests of species co-occurrence and the newly designed approaches for the permutation test with the constraints on spatial autocorrelation and species association, we here present an R package 'SpatialNullModel'. This package contains all the functions that allows the permutation test to be carried out while keeping the independence of species in their decision to colonise sites constant during the randomization of the observed data to generate a sampling distribution. In particular, it presents the functions that keep the spatial autocorrelation of species using the Moran's I coefficient and the association between species using the "Association Index" (proposed by Dice (1945)) constant during the permutation test. The following are the topics documented with their brief descriptions. The details on their usage are presented in the Appendix section.

4.1.1 Spatial1

This represents the co-occurrence simulation algorithm used in generating the sampling distribution to be used in the permutation test. The algorithm works by fixing the level of independence of species in the site-by-species matrix during the randomization procedures to generate random samples. The level of independence represented as spatial autocorrelation is quantified using the Moran's I coefficient. The Moran's I value of the random samples generated has been computed such that this value is approximately the same as the Moran's I value of the observed data.

4.1.2 Spatial2

Like Spatial1, this is a co-occurrence simulation algorithm used to generate a sampling distribution. The algorithm works by fixing the level of independence

of species in the site-by-species matrix during the randomization procedures to generate random samples by keeping constant the ecological association between species pairs. That is, the 'Association Index' value of the observed data is kept constant during randomization so that random samples generated have the same value of the 'Association Index' as the observed data.

4.1.3 SpatialNullModel

'SpatialNullModel' is a collection of functions for calculating the simulation algorithms and community metrics for randomizing the site-by-species data for the spatially explicit null model analysis. It is the engine behind the null model analysis with spatial autocorrelation and species association incorporated.

4.1.4 nullmod2

This is the underlying engine that takes in the site-by-species data and returns the observed data, simulated data (produced using spatial2 algorithm) and C-scores of both the observed data and simulated data.

4.1.5 nullmod1

This function takes in both the site-by-species data and the element of spatial weights matrix which indexes one location relative to the other, and returns the observed data, simulated data (produced using spatial1 algorithm) and C-scores of both the observed and simulated data.

4.1.6 summary.nullmod2

The 'summary.nullmod2' function generates the summary statistics from which the p value result can be inferred and conclusion made on the hypothesis in question. That is, whether the null hypothesis should be accepted or rejected. The algorithm used to generate random samples is "Spatial2".

4.1.7 summary.nullmod1

Like 'summary.nullmod2', this function generates the summary statistics from which the p value result can be inferred and conclusion made on the hypothesis in question. The algorithm used to generate random samples is "Spatial1".

The following sections presents the data and the results obtained when spatially explicit null model was used to test the null hypothesis that species competition plays no role in structuring ecological communities against the alternative hypothesis that the patterns exhibited by the ecological communities

are not by random chance but due to inter-specific competition. The same null hypothesis was tested using the implicit null models. Both results from the two groups of models are compared and conclusion drawn on the effectiveness and accuracy of the models.

4.2 Data used (Caribean Islands)

Caribbean islands are located to the east of Central America and Mexico, southeast of the Gulf of Mexico, and to the north of South America. Born without flora and fauna, they formed animal and plant populations in many ways (O'Keefe, 2016). Some plant species' seeds (for example, mangroves) made a landfall and sprouted after floating in the ocean for months. While other animal and plant species spread by having the former and latter's seed rafted to an island (O'Keefe, 2016).

While some islands have a modicum of animal life, most of what can be observed are insects, birds and lizards (O'Keefe, 2016). This means there are no threatening animals except some dangerous snakes in a few islands. This therefore means Caribbean islands are ideal places for birding, which explains the choice of the study area. Figure 4.1 presents a visualization of the Caribbean islands used in the study, while Table 4.1 gives their specific locations.

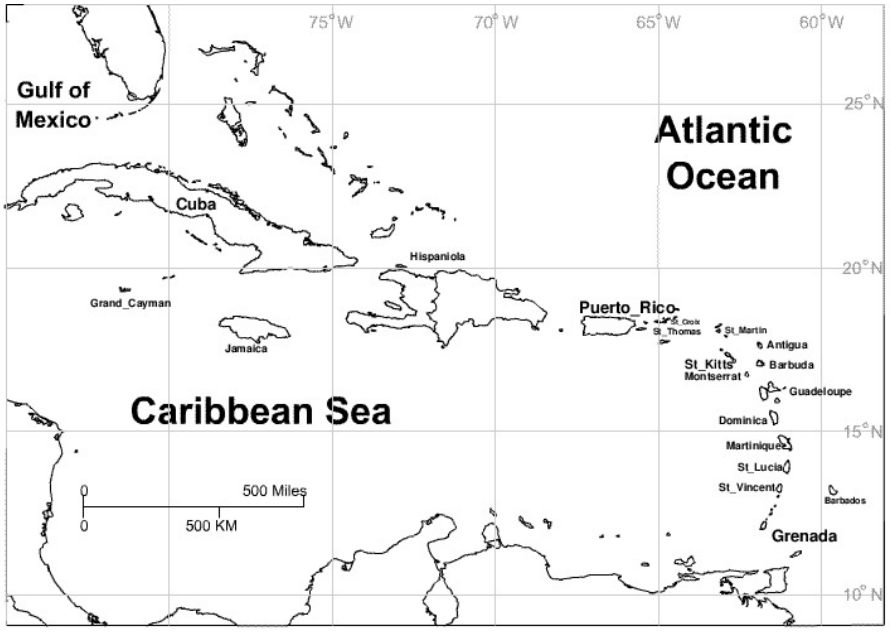


Figure 4.1: Caribbean Islands

Table 4.1: Locations of the study area

	siteNames	Longitudes	Latitudes
1	Cuba	-77.7812	21.5218
2	Hispaniola	-71.5724	19.0019
3	Jamaica	-77.2975	18.1096
4	Puerto_Rico	-66.5901	18.2208
5	Guadeloupe	-61.551	16.265
6	Martinique	-61.0242	14.6415
7	Dominica	-61.371	15.415
8	St_Lucia	-60.9789	13.9094
9	Barbados	-59.5432	13.1939
10	St_Vincent	-61.1863	13.251
11	Grenada	-61.679	12.1165
12	Antigua	-61.8175	17.0747
13	St_Croix	-64.8348	17.7246
14	Grand_Cayman	-81.2409	19.3222
15	St_Kitts	-62.783	17.3578
16	Barbuda	-61.7707	17.6268
17	Montserrat	-62.1874	16.7425
18	St_Martin	-63.0501	18.0708
19	St_Thomas	-64.8941	18.3381

4.3 Spatially explicit and implicit null models compared: Results

To assess the performance of the spatially explicit null model against its spatially implicit counterpart, real dataset from an experiment which was done in the Caribbean islands (refer to Figure 4.1 and Table 4.1) is used to test the effect of inter-specific competition in structuring avifauna/bird communities. To do this using the SpatialNullModel R package, the EcoSimR package is first installed and the following steps followed:

```
library("SpatialNullModel")
library("EcoSimR")
load("data.R")
summary.nullmod2(Data)
summary.coocnullmod(cooc_null_model(dataWiFinches, algo="sim9",nReps=1000))
```

The summary statistics from which the p value result can be inferred using "spatial2" and "spatial1" randomization algorithms are produced by running

the commands: "summary.nullmod2(Data)" and "summary.nullmod1(Data)" respectively in R console. The outputs of these two functions are:

```
Algorithm: Spatial2
Observed C-score:  3.7941
Mean Of Simulated C-scores:  2.9298
Variance Of Simulated C-scores:  0.08374
L.C.V-Lower 95% (1-tail):  2.4482
U.C.V-Upper 95% (1-tail):  3.3824
L.C.V-Lower 95% (2-tail):  2.3528
U.C.V-Upper 95% (2-tail):  3.4559
P-value = 0.001
```

and

```
Algorithm: spatial1
Observed C-score:  3.7941
Mean Of Simulated C-score:  4.0117
Variance Of Simulated C-score:  0.010934
L.C.V-Lower 95% (1-tail):  3.8971
U.C.V-Upper 95% (1-tail):  4.2426
U.C.V-Lower 95% (2-tail):  3.8824
U.C.V-Upper 95% (2-tail):  4.2794
P-value = 0.001
```

respectively. Figures 4.2 and 4.3 illustrates the confidence intervals (marked by blue lines) within which the null hypothesis should be accepted, when spatial2 and spatial1 algorithms were used respectively to generate the random samples.

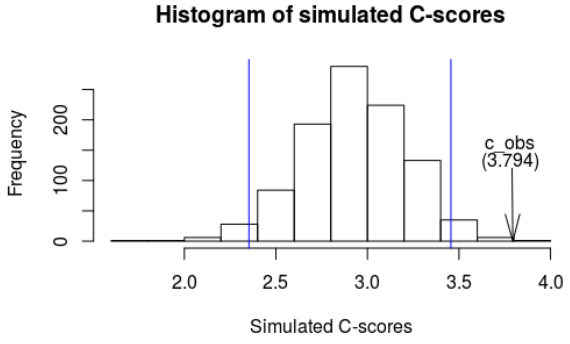


Figure 4.2: A visualization of an histogram illustrating the 2.5% and 97.5% percentiles of the simulated C-scores generated using spatial2 algorithm. The critical values, marked by vertical blue lines, form a confidence interval within which the null hypothesis should be accepted. As illustrated, the C-score value of the observed data (labelled 'c_obs') is outside this interval, implying the null hypothesis should be rejected.

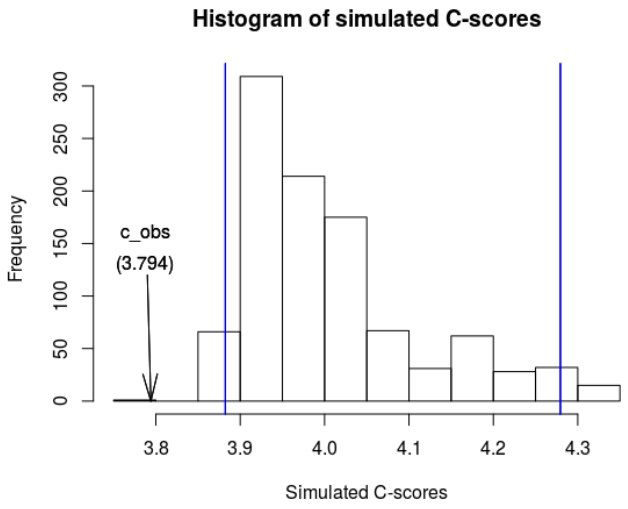


Figure 4.3: A visualization of an histogram illustrating the 2.5% and 97.5% percentiles of the simulated C-scores generated using spatial1 algorithm. Like Figure 4.2, the blue vertical lines mark the critical values which form a confidence interval within which the null hypothesis should be accepted. As illustrated, the C-score value of the observed data (labelled 'c_obs') is also outside this interval, implying the null hypothesis should be rejected.

Similarly, part of the summary statistics produced when "sim9" algorithm of the spatially implicit null model was used are:

```
Metric:  c_score
Algorithm:  sim9
Observed Index:  3.7941
Mean Of Simulated Index:  2.8166
Variance Of Simulated Index:  0.038085
Lower 95% (1-tail):  2.5147
Upper 95% (1-tail):  3.1912
Lower 95% (2-tail):  2.4998
Upper 95% (2-tail):  3.2941
Lower-tail P >  0.999
Upper-tail P <  0.001
Observed metric > 1000 simulated metrics
Observed metric < 0 simulated metrics
Observed metric = 0 simulated metrics
```

Part of summary statistics for the rest of the algorithms are presented in Table 4.2. Figure 4.4 illustrates the confidence intervals when sim1 to sim9 algorithms were used to generate the random samples for both one-tail and two-tailed tests. The intervals are marked by long-dashed and short-dashed vertical lines for both one-tailed and two-tailed tests respectively (Gotelli, 2000). These intervals mark the regions within which the null hypothesis should be

accepted. These summary statistics are then compared with regard to their p values or critical values and consequently the inferences are drawn from the two groups of results- those of the spatially explicit null models and the classic (spatially implicit) null models. Ideally, the inference made from the outputs of a null model observing all the conditions of the test including; observance of the assumptions of the hypothesis test such as independence and considering the biasses (making the null model overly conservative) generated during the randomization procedures to generate random samples using either of the two groups of null models, should be preferred.

Table 4.2: Summary of the results

		Simulated C-scores			
		One-tailed test		**Two-tailed test**	
Algorithms	**Observed**	**Lower**	**Upper**	**Lower**	**Upper**
Sim1	3.7941	6.3529	8.2336	6.2278	8.5751
Sim2	3.7941	2.4335	3.3603	2.3162	3.4412
Sim3	3.7941	6.6173	8.1905	6.4926	8.6095
Sim4	3.7941	2.0290	3.0662	1.9408	3.1767
Sim5	3.7941	4.2247	7.5477	3.9191	8.0011
Sim6	3.7941	5.5147	7.6860	5.3158	7.8925
Sim7	3.7941	4.0509	7.7571	3.8129	8.1541
Sim8	3.7941	3.6286	7.0000	3.4230	7.4872
Sim9	3.7941	2.5147	3.1912	2.4998	3.2941
spatial1	3.7941	3.8971	4.2426	3.8824	4.2794
spatial2	3.7941	2.4482	3.3824	2.3528	3.4559

Notes: Observed C-scores and the lower and upper critical values of the C-scores of the simulated matrices for both the one-tailed and two-tailed tests for the 9 spatially implicit and 2 spatially explicit randomization algorithms.

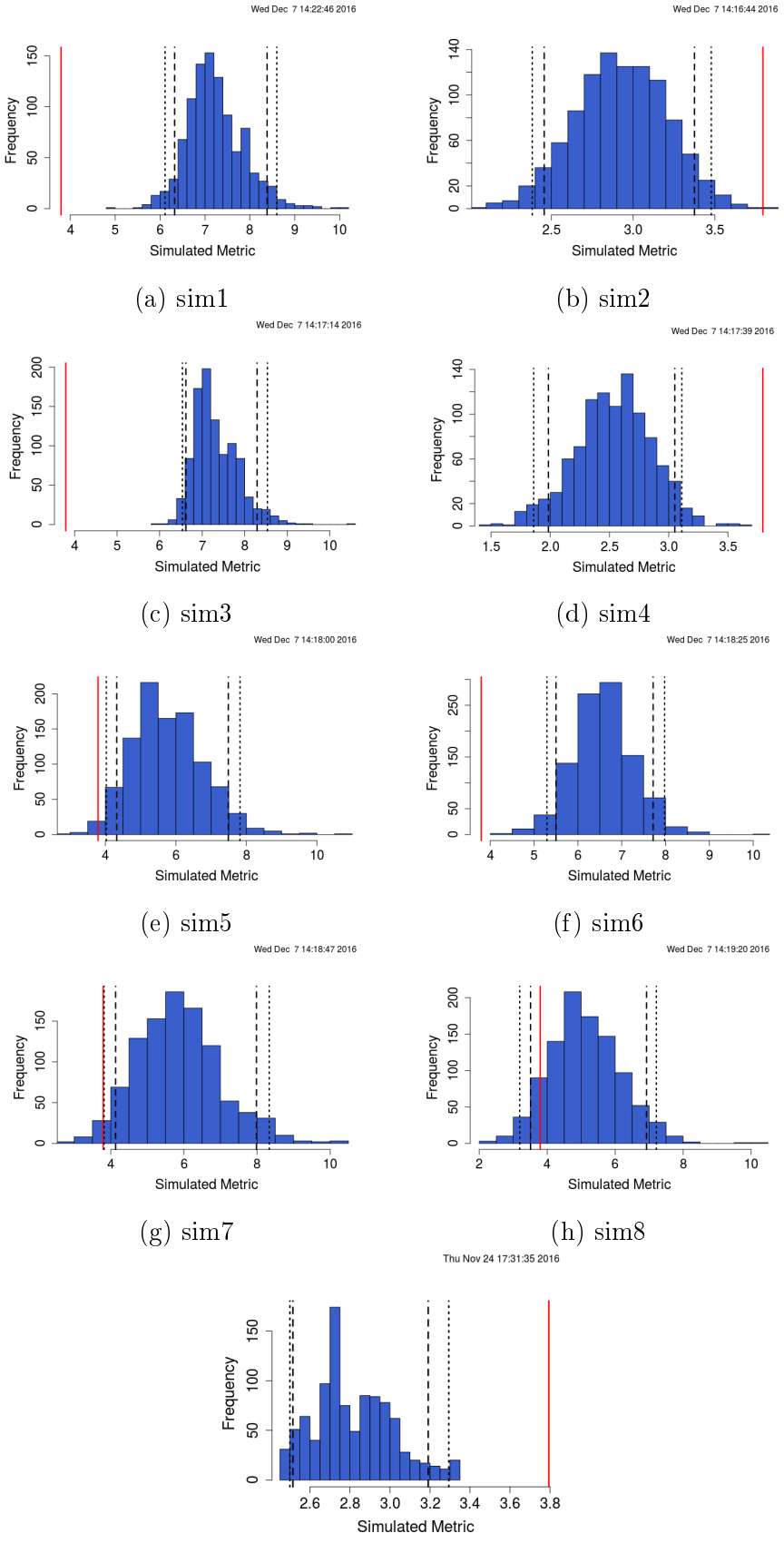


Figure 4.4: A visualization of the histograms illustrating the 2.5% and 97.5% percentiles of both one-tailed (marked by long-dashed vertical lines) and two-tailed (marked by short-dashed vertical lines) tests. These critical values form the confidence intervals within which the null hypothesis should be accepted for both one-tailed and two-tailed tests. The C-score value of the observed data is marked by vertical red lines, which is outside the confidence interval for both tests, implying the null hypothesis should be rejected.

4.4 Discussion

From our analysis, the samples from the two spatial null models are generated randomly. For instance, spatial1 algorithm randomizes the observed site-by-species matrix with no constraint imposed and then selects the random samples generated that have approximately the same spatial autocorrelation as the observed matrix. This prevents the null model from being overly conservative. Also, by selecting the random samples with the same spatial autocorrelation as the observed matrix to form a sampling distribution, the model ensures that the test's assumption of independence is not violated. The same procedure is followed with regard to spatial2 algorithm. However, the classic null model algorithm (like sim9) do not account for the spatial autocorrelation of species distributions and the ecological association. This in turn leads to violation of the independence assumption of the test, leading to counterintuitive results by the test. It therefore suffices to conclude that spatially explicit null models are superior to their spatially implicit counterparts.

4.4.1 Statistical & Ecological Interpretation of the results

From the statistical results in section 4.3, ecological understanding can be reached. Figures 4.2, 4.3 and 4.4 and Table 4.2, implies that the null hypothesis should be rejected (except in sim8, Figure 4.4), since the observed C-score value is outside the confidence intervals created by the critical values (in all the algorithms except sim8). Statistically, this means that, at 5% level of significance, we are 95% confident that the null hypothesis is false. On the other hand, this can be interpreted ecologically to mean ecological communities are structured by inter-specific competition, and the patterns observed are therefore not by random chance. Though the results of all the algorithms (except sim8) leads to same conclusion regarding the structure of ecological communities, spatial1 and spatial2

algorithms are more reliable since they don't violate any of the assumptions of the hypothesis test.

Chapter 5

Conclusion

This project has presented the spatially structured null models in ecology. Two problems extant in the spatially implicit null models have been addressed:

i) Non-observance of the statistical independence assumption.

ii) Randomization procedures making the null model overly conservative.

In particular, the often ignored spatial autocorrelation of species distributions (Hui *et al.*, 2010) in a permutation test which could lead to counterintuitive results in the null model test (Hausdorf and Hennig, 2007) has been accounted for by fixing the spatial autocorrelation of species distributions during the randomization of the observed data to generate a sampling distribution. Consequently, like Fuller and Enquist (2012), the test has been made to account for the spatial autocorrelation of each species. Another important thing that has always been ignored in the classic permutation test is the matching of environmental heterogeneity and species' habitat preference. To tease apart the role of environmental heterogeneity from biotic interactions, the permutation test has been allowed to reserve the association between species by fixing the amount of ecological association between species during the randomization of the observed data to generate a sampling distribution.

This project has thus designed a permutation null model test that can progressively include the spatial autocorrelation of species

distributions and the association between species so that the role of aggregation and environmental heterogeneity can be further examined. A R package has been build to implement both classic (spatially implicit) null model tests of co-occurrence and newly designed approaches (spatially explicit null model test) for the permutation test with constraints on species autocorrelation and association. West Indian Finches real dataset has been used for model evaluation.

The results of this study confirmed Diamond (1975)'s hypothesis to be true. That is, ecological communities are structured by inter-specific competition. With reliable results from the newly designed spatially explicit null model, the forces behind the structure of ecological communities are now understood for better conservation and stability of the ecosystems. Understanding the biotic factors making two or more species to co-exist in the same site is paramount in avoiding their extinction and promoting better conservation strategies.

5.1 Recommendations for Further Research

Species aggregation and environmental heterogeneity can determine the structure of ecological communities, and not necessarily inter-species competition. To draw a line between the effects of species aggregation, environmental heterogeneity and inter-specific competition on the structure of ecological communities, further research needs to be done.

Our work concentrated on incorporating, separately, the spatial autocorrelation and ecologic association between species pairs. However, though computationally expensive, the model can be modified to incorporate both ecologic association and spatial autocorrelation of species, simultaneously. This can be done by fixing both spatial autocorrelation and the ecologic association between species

concurrently during randomization of the observed data to generate a sampling distribution. This way, sensitivity analysis can be carried out with these parameters (spatial autocorrelation, species association and both) varied to determine the combination of parameters which produces a robust model.

Appendices

Appendix A

SpatialNullModel Manual

Package 'SpatialNullModel'

December 5, 2016

Type Package

Title Spatially explicit null models

Version 0.1.0

Date 2016-12-05

Author Vitalis Lagat, Cang Hui

Maintainer Vitalis Lagat <vitalis@aims.ac.za>

Description Functions that incorporate the spatial autocorrelation and association in the statistical null model test of species co-occurrence.

License GPL-3

LazyData TRUE

RoxygenNote 5.0.1

R topics documented:

Ass.index *Association index metric*

Description

A metric to quantify the amount of ecological association between species pairs. This function sums all the association indices for all the species pairs and divides by the total number of species pairs. I.e., it computes the average association index per species pair.

Usage

```
Ass.index(Data)
```

Arguments

Data	A site-by-species matrix with entries 1 (indicating the occurrence of a species in a site) and 0 (indicating the absence of species in a site).

Value

Average association index for the entire species colonisation pattern.

References

Dice, L.R.: Measures of the amount of ecological association between species. Ecology, vol. 26, no. 3, pp. 297-302, 1945.

nullmod1 *null model engine for spatial autocorrelation*

Description

This function depends on spatial1 function. It takes in the site-by-species data with the spatial weights matrix to produce c-score of the observed data, c-score of the simulated data, simulated data and real/observed site-by-species data.

Usage

```
nullmod1(Data,weights)
```

Arguments

Data	A site-by-species matrix with entries 1 (indicating the occurrence of a species in a site) and 0 (indicating the absence of a species in a given site).
weights	The distance weights matrix indexing the location of i relative to j.

Value

C-score of the observed data, c-score of the simulated data, real/observed data and simulated data.

`nullmod2` *null model engine for species association*

Description

This function depends on spatial2 function. It takes in the site-by-species data to produce the c-score of the observed data, c-score of the simulated data, simulated data and real/observed site-by-species data.

Usage

```
nullmod2(Data)
```

Arguments

`Data`	A site-by-species matrix with entries 1 (indicating the occurrence of a species in a site) and 0 (indicating the absence of a species in a given site).

Value

C-score of the observed data, c-score of the simulated data, real/observed data and simulated data.

`perm1` *Spatial1 Permutation test*

Description

For every new matrix generated using the spatial1 randomization algorithm, the C-score value is computed. This function gives the distribution of the C-score values from the n simulations performed.

Usage

```
perm1(Data)
```

Arguments

`Data`	A site-by-species matrix with entries 1 (indicating the occurrence of a species in a site) and 0 (indicating the absence of a species in a given site).

Value

Sampling distribution of the C-score values of the random sample matrices.

perm2 *Spatial2 Permutation test*

Description

For every new matrix generated using the spatial2 algorithm, the C-score value is computed. This function gives the distribution of the C-score values from the n simulations performed.

Usage

```
perm2(Data)
```

Arguments

Data	A site-by-species matrix with entries 1 (indicating the occurrence of a species in a site) and 0 (indicating the absence of a species in a given site).

Value

Sampling distribution of the C-score values of the random sample matrices.

sample_data *Data sampling metric*

Description

A metric that generates a sample matrix by randomizing every column of the observed matrix. That is, it is a function that randomizes every column of the observed matrix and replaces the observed columns with the new samples to form a random site-by-species matrix.

Usage

```
sample_data(Data)
```

Arguments

Data	A site-by-species matrix with entries 1 (indicating the occurrence of a species in a site) and 0 (indicating the absence of species in a given site).

Value

A new random site-by-species matrix.

spatial1 *Spatial1 randomization algorithm*

Description

This function performs N simulations and outputs the sample matrix whose Moran's I value is approximately equal to the Moran's I value of the observed matrix.

Usage

```
Spatial1(Data,weights)
```

Arguments

`Data`	A site-by-species matrix with entries 1 (indicating the occurrence of a species in a site) and 0 (indicating the absence of a species in a site).
`weights`	Distance weights matrix indexing the location i relative to j.

Value

A new random site-by-species matrix.

spatial2 *Spatial2 randomization algorithm*

Description

This function performs N simulations and outputs the sample matrix whose Association Index value is equal to the Association Index of the observed matrix.

Usage

```
Spatial2(Data, N, e)
```

Arguments

`Data`	A site-by-species matrix with entries 1 (indicating the occurrence of a species in a site) and 0 (indicating the absence of a species in a site).
`N`	Number of simulations to be performed.
`e`	A small number epsilon which is the difference between the Association Index of the observed matrix and that of the sample matrix. If this value holds true for any given sample relative to the observed data, the species in the sample is deemed to have the 'same' amount of ecological association as the species in the observed data.

Value

A new random site-by-species matrix

`summary.nullmod1` *summary statistics for the null model engine for spatial autocorrelation*

Description

This function computes the summary statistics obtained when spatial1 algorithm is used to generate random samples. It takes as input a null model object (nullmod1) and outputs the summary statistics including the p-value and the critical values forming the confidence interval within which the null hypothesis should be accepted.

Usage

```
summary.nullmod1(nullmodObj1)
```

Arguments

`nullmodObj1`	A null model object

Value

C-score of the observed data, mean and variance of the c-scores of simulated data, p-value and the critical values of both the one-tailed and two-tailed tests.

`summary.nullmod2` *summary statistics for the null model engine for species association*

Description

This function computes the summary statistics obtained when spatial2 algorithm is used to generate random samples. It takes as input a null model object (nullmod2) and outputs the summary statistics including the p-value and the critical values forming the confidence interval within which the null hypothesis should be accepted.

Usage

```
summary.nullmod2(nullmodObj)
```

Arguments

`nullmodObj2`	A null model object

Value

C-score of the observed data, mean and variance of the c-score of simulated data, p-value and the critical values of both the one-tailed and two-tailed tests.

`S_ij` *Number of sites where both species i and j co-occurred*

Description

This function counts the total number of sites harbouring both species i and j.

Usage

```
S_ij(species_i,species_j)
```

Arguments

`species_i`	Species co-occurring with another in the same site
`species_j`	Species whose co-occurrences with another is to be determined

Value

The total number of sites where both species i and j co-occurred

Index

List of References

Ackerly, D., Schwilk, D. and Webb, C. (2006). Niche evolution and adaptive radiation: testing the order of trait divergence. *Ecology*, vol. 87, no. sp7.

Adams, D.C. (2007). Organization of plethodon salamander communities: Guild-based community assembly. *Ecology*, vol. 88, no. 5, pp. 1292–1299.

Bahn, V., J O'Connor, R. and B Krohn, W. (2006). Importance of spatial autocorrelation in modeling bird distributions at a continental scale. *Ecography*, vol. 29, no. 6, pp. 835–844.

Bascompte, J. and Melián, C.J. (2005). Simple trophic modules for complex food webs. *Ecology*, vol. 86, no. 11, pp. 2868–2873.

Betts, M.G., Diamond, A., Forbes, G., Villard, M.-A. and Gunn, J. (2006). The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. *Ecological Modelling*, vol. 191, no. 2, pp. 197–224.

Blüthgen, N., Fründ, J., Vázquez, D.P. and Menzel, F. (2008). What do interaction network metrics tell us about specialization and biological traits. *Ecology*, vol. 89, no. 12, pp. 3387–3399.

Burns, K. and Zotz, G. (2010). A hierarchical framework for investigating epiphyte assemblages: networks, meta-communities, and scale. *Ecology*, vol. 91, no. 2, pp. 377–385.

Coleman, B.D., Mares, M.A., Willig, M.R. and Hsieh, Y.-H. (1982). Randomness, area, and species richness. *Ecology*, vol. 63, no. 4, pp. 1121–1133.

Connor, E.F., Collins, M.D. and Simberloff, D. (2013). The checkered history of checkerboard distributions. *Ecology*, vol. 94, no. 11, pp. 2403–2414.

Connor, E.F. and Simberloff, D. (1979). The assembly of species communities: chance or competition? *Ecology*, vol. 60, no. 6, pp. 1132–1140.

Connor, E.F. and Simberloff, D. (1983). Interspecific competition and species co-occurrence patterns on islands: null models and the evaluation of evidence. *OIKOS*, vol. 41, no. 3, pp. 455–465.

Cornwell, W.K., Schwilk, D.W. and Ackerly, D.D. (2006). A trait-based test for habitat filtering: convex hull volume. *Ecology*, vol. 87, no. 6, pp. 1465–1471.

Diamond, J. (1975). Assembly of species communities in: 'ecology and evolution of communities' (eds. m.l. cody and j.m. diamond) pp. 342–444.

Diamond, J.M. and Gilpin, M.E. (1982). Examination of the "null" model of connor and simberloff for species co-occurrences on islands. *Oecologia*, vol. 52, no. 1, pp. 64–74.

Dice, L.R. (1945). Measures of the amount of ecologic association between species. *Ecology*, vol. 26, no. 3, pp. 297–302.

Diniz-Filho, J.A.F., Bini, L.M. and Hawkins, B.A. (2003). Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography*, vol. 12, no. 1, pp. 53–64.

F Dormann, C., M McPherson, J., B Araújo, M., Bivand, R., Bolliger, J., Carl, G., G Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W. *et al.* (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, vol. 30, no. 5, pp. 609–628.

Fortin, M.-J. and Dale, M.R. (2005). *Spatial analysis: a guide for ecologists*. Cambridge University Press.

Fuller, M.M. and Enquist, B.J. (2012). Accounting for spatial autocorrelation in null models of tree species association. *Ecography*, vol. 35, no. 6, pp. 510–518.

Gilpin, M.E. and Diamond, J.M. (1982). Factors contributing to non-randomness in species co-occurrences on islands. *Oecologia*, vol. 52, no. 1, pp. 75–84.

GIS (2016). Spatial autocorrelation and moran's i in gis. Date accessed: 05 October 2016.
Available at: http://gisgeography.com/spatial-autocorrelation-moran-i-gis/

GITTA (2016). The join count statistic (at a nominal level). Date accessed: 27 July 2016.
Available at: http://www.gitta.info/DiscrSpatVari/en/html/spat_depend_join_ct_stat.html

Gotelli, N.J. (2000). Null model analysis of species co-occurrence patterns. *Ecology*, vol. 81, no. 9, pp. 2606–2621.

Gotelli, N.J. (2001). Research frontiers in null model analysis. *Global ecology and biogeography*, vol. 10, no. 4, pp. 337–343.

Gotelli, N.J. and Abele, L.G. (1982). Statistical distributions of west indian land bird families. *Journal of Biogeography*, vol. 9, no. 42, pp. 421–435.

Gotelli, N.J. and McCabe, D.J. (2002). Species co-occurrence: a meta-analysis of jm diamond's assembly rules model. *Ecology*, vol. 83, no. 8, pp. 2091–2096.

Gotelli, N.J. and Ulrich, W. (2012). Statistical challenges in null model analysis. *Oikos*, vol. 121, no. 2, pp. 171–180.

Gotelli, N.J.G. *et al.* (1996). *Null models in ecology*. 574.501519 G6. Smithsonian Institution Press.

Griffith, D.A. and Chun, Y. (2015). Spatial autocorrelation in spatial interactions models: geographic scale and resolution implications for network resilience and vulnerability. *Networks and Spatial Economics*, vol. 15, no. 2, pp. 337–365.

Hausdorf, B. and Hennig, C. (2007). Null model tests of clustering of species, negative co-occurrence patterns and nestedness in meta-communities. *Oikos*, vol. 116, no. 5, pp. 818–828.

Helmus, M.R., Savage, K., Diebel, M.W., Maxted, J.T. and Ives, A.R. (2007). Separating the determinants of phylogenetic community structure. *Ecology Letters*, vol. 10, no. 10, pp. 917–925.

Hui, C. (2015). Unlocking patterns of nature-the marriage of mathematics and ecology. Stellenbosch: Stellenbosch University, 2015.

Hui, C., Veldtman, R. and McGeoch, M.A. (2010). Measures, perceptions and scaling patterns of aggregated species distributions. *Ecography*, vol. 33, no. 1, pp. 95–102.

Ingram, T. and Shurin, J.B. (2009). Trait-based assembly and phylogenetic structure in northeast pacific rockfish assemblages. *Ecology*, vol. 90, no. 9, pp. 2444–2453.

Kembel, S.W. (2009). Disentangling niche and neutral influences on community assembly: assessing the performance of community phylogenetic structure tests. *Ecology Letters*, vol. 12, no. 9, pp. 949–960.

Kesey-Bear (2016). Co-occurrence. Date accessed: 5 December 2016. Available at: http://www.esapubs.org/archive/ecol/E081/022/EcoSim%20Help/CoOccur/CoOccurrence.htm

Krasnov, B.R., Stanko, M. and Morand, S. (2006). Are ectoparasite communities structured? species co-occurrence, temporal variation and null models. *Journal of Animal Ecology*, vol. 75, no. 6, pp. 1330–1339.

Lay, S. (2016). What is competition? Date accessed: 05 December 2016. Available at: https://infogr.am/what-is-competition

Lichstein, J.W., Simons, T.R., Shriner, S.A. and Franzreb, K.E. (2002). Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, vol. 72, no. 3, pp. 445–463.

Liu, Y., Tong, D. and Liu, X. (2015). Measuring spatial autocorrelation of vectors. *Geographical Analysis*, vol. 47, no. 3, pp. 300–319.

Mathiba, M. and Awuah-Offei, K. (2015). Spatial autocorrelation of soil co2 fluxes on reclaimed mine land. *Environmental Earth Sciences*, vol. 73, no. 12, pp. 8287–8297.

MBASKOOL (2016). Null model. Date accessed: 7 October 2016. Available at: http://www.mbaskool.com/business-concepts/statistics/7520-null-model.html

McCoy, E.D. and Heck, K. (1987). Some observations on the use of taxonomic similarity in large-scale biogeography. *Journal of Biogeography*, vol. 14, pp. 79–87.

Melecky, L. (2015). Spatial autocorrelation method for local analysis of the eu. *Procedia Economics and Finance*, vol. 23, pp. 1102–1109.

Moran, P.A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, vol. 37, no. 1/2, pp. 17–23.

Mouillot, D., Krasnov, B.R. and Poulin, R. (2008). High intervality explained by phylogenetic constraints in host–parasite webs. *Ecology*, vol. 89, no. 7, pp. 2043–2051.

Nicholas J. Gotelli, Edmund M. Hart, A.M.E. (2016). Co-occurrence analysis. Date accessed: 12 October 2016. Available at: ftp://cran.r-project.org/pub/R/web/packages/EcoSimR/vignettes/CoOccurrenceVignette.html

O'Keefe, M.T. (2016). Caribbean flora & fauna. Date accessed: 05 December 2016. Available at: http://guidetocaribbeanvacations.com/flora_fauna/

Pielou, D. and Pielou, E. (1968). Association among species of infrequent occurrence: the insect and spider fauna of polyporus betulinus (bulliard) fries. *Journal of Theoretical Biology*, vol. 21, no. 2, pp. 202–216.

PlantLife (2016). Competition. Date accessed: 05 December 2016. Available at: http://lifeofplant.blogspot.co.za/2011/05/competition.html

Robson, D. (1972). Appendix: statistical tests of significance. *Journal of Theoretical Biology*, vol. 34, pp. 350–352.

Sanderson, J.G., Diamond, J.M. and Pimm, S.L. (2009). Pairwise co-existence of bismarck and solomon landbird species. *Evolutionary Ecology Research*, vol. 11, no. 5, pp. 771–786.

Schluter, D. (1984). A variance test for detecting species associations, with some example applications. *Ecology*, vol. 65, no. 3, pp. 998–1005.

Sokal, R. and Rohlf, F. (1995). Biometry: the principles and practice of statistics in biological research. *New York*.

Stone, L. and Roberts, A. (1990). The checkerboard score and species distributions. *Oecologia*, vol. 85, no. 1, pp. 74–79.

Tobler, W.R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, vol. 46, no. sup1, pp. 234–240.

Veech, J.A. (2012). Significance testing in ecological null models. *Theoretical ecology*, vol. 5, no. 4, pp. 611–616.

Westerholt, R., Resch, B. and Zipf, A. (2015). A local scale-sensitive indicator of spatial autocorrelation for assessing high-and low-value clusters in multiscale datasets. *International Journal of Geographical Information Science*, vol. 29, no. 5, pp. 868–887.

Wikipedia (2016). Species distribution. Date accessed: 25 July 2016. Available at: https://en.wikipedia.org/wiki/Species_distribution

Zierahn, U. (2012). The importance of spatial autocorrelation for regional employment growth in germany. *Jahrbuch für Regionalwissenschaft*, vol. 32, no. 1, pp. 19–43.

Zimmermann, Y., Ramírez, S.R. and Eltz, T. (2009). Chemical niche differentiation among sympatric species of orchid bees. *Ecology*, vol. 90, no. 11, pp. 2994–3008.